An Augmented Reality Human-Robot Collaboration System

Scott A. Green

A thesis
submitted in partial fulfilment
of the requirements for the degree
of
Doctor of Philosophy
in
Mechanical Engineering
at the
University of Canterbury,
Christchurch, New Zealand.

2008

Dedicated to:

Cecil

May he rest in peace

and

Annie

Abstract

Although robotics is well established as a research field, there has been relatively little work on human-robot collaboration. This type of collaboration is going to become an increasingly important issue as robots work ever more closely with humans. Clearly, there is a growing need for research on human-robot collaboration and communication between humans and robotic systems.

Research into human-human communication can be used as a starting point in developing a robust human-robot collaboration system. Previous research into collaborative efforts with humans has shown that grounding, situational awareness, a common frame of reference and spatial referencing are vital in effective communication. Therefore, these items comprise a list of required attributes of an effective human-robot collaborative system.

Augmented Reality (AR) is a technology for overlaying three-dimensional virtual graphics onto the user's view of the real world. It also allows for real time interaction with these virtual graphics, enabling a user to reach into the augmented world and manipulate it directly. The internal state of a robot and its intended actions can be displayed through the virtual imagery in the AR environment. Therefore, AR can bridge the divide between human and robotic systems and enable effective human-robot collaboration.

This thesis describes the work involved in developing the Augmented Reality Human-Robot Collaboration (AR-HRC) System. It first garners design criteria for the system from a review of communication and collaboration in human-human interaction, the current state of Human-Robot Interaction (HRI) and related work in AR. A review of research in multimodal interfaces is then provided highlighting the benefits of using such an interface design. Therefore, an AR multimodal interface was developed to determine if this type of design improved performance over a single modality design. Indeed, the multimodal interface was found to improve performance, thereby providing the impetus to use a multimodal design approach for the AR-HRC system.

The architectural design of the system is then presented. A user study conducted to determine what kind of interaction people would use when collaborating with a mobile robot is discussed and then the integration of a mobile robot is described. Finally, an evaluation of the AR-HRC system is presented.

Contents

1 Introduction **1**
- 1.1 Chapter Summary . . . 3
- 1.2 Acknowledgments . . . 4

2 Review: Human-Robot Interaction **5**
- 2.1 Communication and Collaboration . . . 5
- 2.2 Human-Robot Interaction . . . 7
  - 2.2.1 Robots as Tools . . . 7
  - 2.2.2 Guide, Hosting and Assistant Robots . . . 9
  - 2.2.3 Humanoid Robots . . . 14
  - 2.2.4 Robots in Collaborative Tasks . . . 17
- 2.3 Summary . . . 20

3 Augmented Reality for Human-Robot Collaboration **21**
- 3.1 Augmented Reality . . . 22
  - 3.1.1 Introduction to Augmented Reality . . . 22
  - 3.1.2 AR in Collaborative Tasks . . . 24
  - 3.1.3 AR in Human-Robot Interaction . . . 32
  - 3.1.4 Summary . . . 34
- 3.2 Multimodal Interaction . . . 36
- 3.3 Wizard of Oz Study . . . 38
- 3.4 Design Guidelines . . . 41
- 3.5 Summary . . . 45

4 Multimodal AR Interaction **47**
- 4.1 Architecture . . . 47
- 4.2 MARS Application . . . 49
- 4.3 The Modalities . . . 51
  - 4.3.1 Gesture Only . . . 52
  - 4.3.2 Speech with Static Paddle Gestures . . . 53
  - 4.3.3 Speech and Gesture . . . 54
- 4.4 Speech Commands . . . 55

4.5 Evaluation . . . 55
4.6 Summary . . . 57

5 Architectural Design **59**
5.1 Design Approach . . . 59
5.1.1 Speech Processing . . . 61
5.1.2 Dialog Management System . . . 63
5.1.3 Gesture Processing . . . 64
5.1.4 Viewpoint Processing . . . 65
5.1.5 HRC-ARE . . . 67
5.1.6 Multimodal Communication Processor . . . 67
5.2 Deeper Spatial Dialog . . . 67
5.3 Summary . . . 73

6 Multimodal Metric Study **75**
6.1 Experimental Design . . . 75
6.1.1 Set Up . . . 76
6.1.2 Experimental Conditions . . . 79
6.1.3 Procedure . . . 79
6.1.4 Participants . . . 84
6.2 Results . . . 84
6.2.1 Objective Measures . . . 85
6.2.2 Pre-Experiment Questionnaire . . . 87
6.2.2.1 Speech Only Condition . . . 87
6.2.2.2 Gesture Only Condition . . . 88
6.2.2.3 Speech and Gesture Condition . . . 88
6.2.2.4 Comparison to Questionnaire with Robot . . . 89
6.2.3 Experimental Results . . . 89
6.2.3.1 Speech Only Condition . . . 89
6.2.3.2 Gesture Only Condition . . . 91
6.2.3.3 Speech and Gesture Condition . . . 92
6.2.4 Post-Experiment Questionnaire . . . 93
6.3 Discussion . . . 95
6.4 Design Guidelines . . . 98
6.5 Summary . . . 98

7 Integration with a Mobile Robot **101**
7.1 Interaction Techniques . . . 101
7.2 Integration . . . 106
7.3 Interaction Scenario . . . 108
7.4 Summary . . . 113

8 System Evaluation **115**
8.1 Experimental Design 115
8.2 Participants 117
8.3 Procedure 117
8.3.1 Immersive Condition 119
8.3.2 Speech and Gesture no Planning 120
8.3.3 Speech and Gesture with Planning, Review and Modification 120
8.4 Results 121
8.4.1 Objective Measures 122
8.4.2 Subjective Measures 125
8.4.3 Participant Comments 128
8.5 Discussion 129
8.6 Summary 131

9 Conclusions **133**

10 Future Work **139**
10.1 AR-HRC Modules 139
10.2 The AR-HRC System 140
10.3 Integration and Evaluation Studies 141

A Wizard of OZ Study Questionnaires **143**

B Interface Evaluation Questionnaires **153**

References **157**

List of Figures

2.1 CMU Host Robot Sage 10
2.2 Robovie 11
2.3 Gestureman 12
2.4 Mel 13
2.5 Cero 13
2.6 Robonaut and Remote Human Operator 15
2.7 Robonaut Interacting in Direct Manner 15
2.8 Kismet 16
2.9 Leonardo 17

3.1 Video See-through AR Interface 23
3.2 AR Video See Through Process 24
3.3 Milgram's Continuum 24
3.4 Shared Space 25
3.5 The MagicBook 26
3.6 AR Remote Collaborator 27
3.7 Mobile AR Setup 28
3.8 AR Interactive Theater Experience 28
3.9 AR Human Pacman 29
3.10 AR Tour Guide 30
3.11 AR Tour Guide Display 30
3.12 Task Space Experiment 31
3.13 AR Interactive Path Planning 32
3.14 AR Path Node Creation 33
3.15 AR Spatial Displays 43

4.1 Handheld Paddle for VOMAR Application 48
4.2 MARS Architecture 49
4.3 MARS Menu Page 50
4.4 MARS Virtual Room 51
4.5 MARS Virtual Object Placement 52
4.6 MARS Speech Process Diagram 54
4.7 MARS Demonstration 56

5.1 AR-HRC System Architecture 62
5.2 Code Snippet: Speech Processing 63
5.3 Code Snippet: Dialog Management System 64
5.4 Code Snippet: Gesture Processing 66
5.5 Code Snippet: Speech Processing II 69
5.6 Code Snippet: Dialog Management System II 69
5.7 Code Snippet: Gesture Processing II 70
5.8 Definition of Spatial Predicate 71
5.9 Code Snippet: Define Behind Position 72

6.1 Wizard Command Center 77
6.2 WOZ Study Video Capture 78
6.3 User Environment 78
6.4 Driving Test Question Example 80
6.5 Example Pre-experiment Questionnaire 80
6.6 Go Around Object 81
6.7 Go Around Identifiable Object 82
6.8 Metric Study Maze 83
6.9 WOZ Completion Times 86
6.10 WOZ Number of Collisions 86
6.11 WOZ Distance Traveled 87
6.12 Gesture Commands Used 92
6.13 WOZ Modality Preference 96
6.14 Multimodal Gesture 97

7.1 Paddle Modality: As Pointer 102

7.2 Icons for Paddle in Gesture Mode 103
7.3 Heads Up Display 106
7.4 The Mobile Robot Integrated into System 107
7.5 Human at Command Center 109
7.6 Interaction Scenario: Go Here 110
7.7 Interaction Scenario: Via Point in Front of Object 111
7.8 Interaction Scenario: Robot Requests Help 112
7.9 Interaction Scenario: Ego and Exo Views 113

8.1 Maze for User Study Task 118
8.2 Performance Study Participant 119
8.3 Performance Study Immersive Condition 120
8.4 Performance Study SGnoP Condition 121
8.5 Performance Study SGwPRM Condition 122
8.6 Performance Study Completion Times 123
8.7 Performance Study Accuracy 124
8.8 Performance Study Close Calls 124
8.9 Performance Study Post Trial Questionnaire 126
8.10 Performance Study Post Experiment Questionnaire 128

List of Tables

4.1 MARS Paddle Commands 53

6.1 Questionnaire Responses Speech Commands 87
6.2 Commands in Speech Only Condition 91
6.3 Post Experiment Questionnaire Answers 95

8.1 Communication Channels for the Virtual Robot 116
8.2 Evaluation Trial Sequences 122

Acknowledgments

I suppose the first person to acknowledge would be J dot Geoffrey Chase. It's actually kinda his fault that I ended up at the University of Canterbury. In my search for graduate schools, Geoff was the only one who responded to my email inquiries with other than a "look at our website" answer. I thought that was a silly response to give to someone who obviously got the email address used for correspondence from said website, especially considering that one of the main concerns I was trying to address was how much support a given university would provide. By answering *all* of my silly questions and encouraging me to apply, Geoff assured me I would be in good hands at Canterbury.

The next tip of the hat would then go to Mark Billinghurst. It took two attempts, but Mark's determination paid off by finding a research project that suited my interests, and thanks again to Geoff for being patient during the time this took to happen. Not only did the topic of this thesis turn out to be a great project to work on over the past four years, but through Mark's contacts I was able to land an internship in sunny California. That internship evolved into full-time employment, for which I am eternally grateful. As is my wife, who rather enjoys the sunshine and career she has since begun.

I would like to thank XiaoQi Chen for his efforts as well. Not only did he continually provide opportunities for publications, but he was nice enough to lug the 5 kg recovered laptop to Las Vegas where we attended a conference together. A conference that he recommended I submit to.

Of course, I have to extend my warmest regards to staff and students of the HIT Lab NZ. Again, kudos to Mark for providing this wonderful environment. Julian, though, deserves special thanks. No matter what silly computer science related question this Mechanical Engineer had, Julian was kind enough to take the time to thoroughly explain the answer. Looking back, it must have seemed pretty odd to have to explain just exactly what "this" means in a C++ class.

I would also like to thank Randy Stiles and Scott Richardson. They brought me in as an intern and then recommended me for full-time employment. For that, I am very grateful.

I would like to thank my parents as well, Cecil and Annie. Unfortunately my father won't be able to see me finish. He supported me in this endeavor, although he did not really understand what I was attempting to achieve. As one who had always looked for a steady job, he was completely taken aback when I explained to him that I intended to resign my position as an Engineer to go back to school. For months my father repeatedly asked me "But it was a good job, right?". Yes Dad, it was a good job. However, by pursuing a PhD I have opened up new doors of opportunity. I miss you Dad.

Finally, I would like to thank my better half, Luz Maria. Throughout this entire ordeal she has given me all her support, encouragement and love. I can not imagine how I would have been able to complete this work without her love and support. Thank you Silly. We both look forward to getting our lives back!

Chapter 1

Introduction

Interface design for Human-Robot Interaction (HRI) will soon become one of the toughest challenges that the field of robotics faces (Thrun, 2004). As HRI interfaces mature it will become more common for humans and robots to work together in a collaborative manner. However, although robotics is well established as a research field, there has been relatively little research on human-robot collaboration (Fong and Nourbakhsh, 2005).

There are many application domains that would benefit from effective human-robot collaboration. For example, in space exploration, recent research has pointed out that to reduce human workload, costs, fatigue driven error and risk, intelligent robotic systems will need to be a significant part of mission design (Fong and Nourbakhsh, 2005). Fong and Nourbakhsh (2005) also observe that scant attention has been paid to joint human-robot teams, and that making human-robot collaboration natural and efficient is crucial to future space exploration. Effective human-robot collaboration will also be required for terrestrial applications such as Urban Search and Rescue (USAR) (Casper and Murphy, 2002; Drury et al., 2005) and tasks completed robotically in hazardous environments, such as the removal of nuclear waste (Tsoukalas and Bargiotas, 1996).

Truly effective collaboration can take place only when the participants are able to communicate with each other in a natural manner. Communicating naturally for humans typically means using a combination of speech, gesture and non-verbal cues such as gaze. Grounding, the common understanding between conversational participants (Clark and Brennan, 1991), shared spatial referencing and spatial awareness are crucial components of communication

and therefore collaboration. In this research, a focus is placed on the development of a human-robot system that is capable of using a range of cues to establish common ground.

In a collaborative team effort, it is important to capitalize on the strengths of each member of the team. For example, humans are good at problem solving and dealing with unexpected events while robots are good at repeated physical tasks and working in hazardous environments. So an effective human-robot collaboration system should exploit these strengths.

One of the key technologies applied in this research on human-robot collaboration is Augmented Reality (AR). AR is a technology for overlaying three-dimensional virtual graphics onto the users view of the real world (Azuma, 1997). AR allows real time interaction with these graphics, enabling a user to reach into the augmented world and manipulate it directly. AR has been shown to be useful for many application areas such as medical (Nikishkov and Tsuchimoto, 2007; Soler et al., 2004), education (Fjeld et al., 2003; Shelton and Hedley, 2002), industry (Friedrich, 2002; Goose et al., 2003), architecture (Sareika and Schmalstieg, 2007) and entertainment (Nilsen et al., 2004; Piekarski and Thomas, 2002).

AR is used to provide a common 3D graphic of the robot's workspace that both the human and robot can reference, and so provide shared spatial understanding. The internal state of the robot and its intended actions are displayed through the virtual imagery in the AR environment. The human team member is thus able to maintain situation awareness of the robot and its surrounding, thereby giving the human-robot team the ability to ground their communication. By coupling AR with spoken dialog a multimodal interface has been developed that enables natural and efficient communication between the human and robot team members, thus enabling effective collaboration.

There is a need for research on different types of HRI systems. This thesis describes the development of the Augmented Reality Human-Robot Collaboration (AR-HRC) system. Fundamentally, this system enables humans to communicate with robotic systems in a natural manner through spoken dialog and gesture interaction, using Augmented Reality technology for visual feedback. This approach is in contrast to the typical reliance on a narrow communication link.

Therefore, the AR-HRC system provides the user the feeling of telepresence. In essence, telepresence is projecting the human into the remote world of the robotic system being collaborated with and providing the means to interact within that environment. The AR-HRC system could have a broad base for usage. For example, robust human-robot interaction enabling collaboration through a shared workspace provided by AR could be applied in remote surveying and exploration by autonomous robotic systems whether these systems are in inhospitable places on earth, in space or on distant moons or planets. Closer to home, this type of system could be used in industrial environments enabling a remote team of humans to collaborate with autonomous robotic systems on the manufacturing floor. The result would be the ability to locate industrial sites away from populated environments, reduce travel to and from these sites thus having a positive impact on the environment from reduced pollution and reduce, if not eliminate, personal injury in these types of industrial settings.

1.1 Chapter Summary

This section provides a road-map of the chapters that make up this thesis.

Chapter 2 Review: Human-Robot Interaction presents related work in human-human communication and collaboration. A survey of Human-Robot Interaction is given that is broken down into robots used as tools, guides, hosting and assistant robots, humanoid robots, and robots in collaborative tasks.

Chapter 3 Augmented Reality for Human-Robot Collaboration introduces Augmented Reality (AR) and provides a survey of AR in human-human collaborative efforts and HRI. Multimodal interaction is introduced and a brief survey is given. The concept of a Wizard of Oz (WOZ) study is presented and the benefits of running such a study are provided. Finally, the results from the related work reviewed in this chapter and Chapter 2 are summarized and an avenue for the development of the Augmented Reality Human-Robot Collaboration system is presented.

Chapter 4 Multimodal AR Interaction presents work completed in the development of a multimodal AR application that provided the impetus

to use a multimodal AR approach in the design of the AR-HRC system.

Chapter 5 Architectural Design describes the architecture of the AR-HRC system.

Chapter 6 Multimodal Metric Study reports on a Wizard of Oz (WOZ) study conducted to define the nature of speech and gestures for human-robot collaboration (HRC). The participants completed a task with a mobile robot using three conditions; speech only, gesture only and speech combined with gesture.

Chapter 7 Integration with a Mobile Robot discusses the integration of a mobile robot into the AR-HRC system and provides the reader with two examples of how interaction with the system takes place.

Chapter 8 System Evaluation reports on a formal evaluation of the AR-HRC system. A typical teleoperation interface is compared to two versions of the AR-HRC system. The full version of the AR-HRC system incorporates spatial dialog, gesture interaction, planning, review and modification of a task plan. A scaled down version of the AR-HRC system is also compared that does not include planning, review or modification. These three interfaces have different communication channels and thus support different types of collaboration.

Chapter 9 Conclusions provides a concise summary of the work completed in this thesis.

Chapter 10 Future Work proposes directions for future research.

1.2 Acknowledgments

The work presented in Chapter 6 was supported by Internal Research and Development (IRAD) funding provided by the Lockheed Martin Space Systems Company Advanced Technology Center, Sunnyvale, CA, USA.

Chapter 2

Review: Human-Robot Interaction

This chapter begins by discussing related work on human-human communication and collaboration, and is followed by a review of the current state of Human-Robot Interaction (HRI). This work is separated into robots as tools, robots as guides, robots as hosts and assistants, humanoid robots, and finally robots in collaborative tasks. The chapter ends by providing a summary of the lessons learned in the review of the current state of HRI, as well as those learned from the review of communication and collaboration. These items form a list of requirements necessary for an effective human-robot collaborative system.

2.1 Communication and Collaboration

In this work, collaboration is defined as "working jointly with others or together especially in an intellectual endeavor" (MerriamWebster, 2008). In addition, Nass et al. (1994) noted that social factors governing human-human interaction equally apply to human-computer interaction. Therefore, before research in human-robot interaction is discussed, human-human communication is briefly reviewed.

There is a vast body of research relating to human-human communication and collaboration. It is clear that people use speech, gesture, gaze and non-verbal cues to attempt to communicate in the clearest possible fashion. In many cases, face-to-face collaboration is also enhanced by, or relies on, real objects or parts of the user's real environment. This section briefly reviews the roles conversational cues and real objects play in face-to-face human-human collaboration. This information is used to provide guidelines for attributes

that robots should have to effectively support human-robot collaboration.

A number of researchers have studied the influence of verbal and non-verbal cues on face-to-face communication. Gaze plays an important role in face-to-face collaboration by providing visual feedback, regulating the flow of conversation, communicating emotions and relationships, and improving concentration by restriction of visual input (Argyle, 1967; Kendon, 1967).

In addition to gaze, humans use a wide range of non-verbal cues to assist in communication, such as nodding (Watanuki et al., 1995) gesture (McNeill, 1992), and posture (Cassell et al., 2001). In many cases, non-verbal cues can only be understood by considering co-occurring speech, such as when using deictic gestures, for example pointing at a location in space and using ambiguous speech (Kendon, 1983). In studying the behavior of human demonstration activities, it was observed that before conversational partners pointed to an object, they always looked in the direction of that object first (Sidner and Lee, 2003).

Real objects and interactions with the real world can also play an important role in collaboration. Minneman and Harrison (1996) show that real objects are more than just a source of information, they are also the constituents of collaborative activity, create reference frames for communication and alter the dynamics of interaction. In general, communication and shared cognition are more robust because of the introduction of shared objects.

Real world objects can be used to provide multiple representations and result in increased shared understanding (Clark and Wilkes-Gibbs, 1986). Fussell et al. (2003) showed through user studies that a shared visual workspace enhances collaboration as it increases awareness. A shared visual workspace also results in fewer verbal exchanges during a collaborative effort.

Clark and Brennan (Clark and Brennan, 1991) provide a communication model to interpret collaboration. In this model, conversation participants attempt to reach shared understanding or common ground. Common ground refers to the set of mutual knowledge, shared beliefs and assumptions that collaborators have. This process of establishing shared understanding, or "grounding", involves communication using a range of modalities including voice, gesture, facial expression and non-verbal body language.

This look into communication and collaboration provides some of the requirements that will be necessary for an effective human-robot collaboration system. The robot will need to be able to recognize and produce non-verbal communication cues to be an effective collaborative partner. The human should also be aware of the robot in its surroundings and the interaction of collaborative partners within those surroundings, in essence maintaining situation awareness (Endsley, 1995). The human-robot team should also be able to communicate effectively and reach common ground easily.

2.2 Human-Robot Interaction

The next four sections review current research in HRI. These sections have been separated into the use of robots as tools, robots as guides and assistants, the development of humanoid robots, and the use of robots in collaborative tasks. Finally, a summary is given discussing the attributes required of an effective human-robot collaborative system based on the current state of research in HRI. In addition, Bekey et al. (2008) provide an excellent overview of the state of art of robotics with a focus on how this research is taking places in various regions around the world.

2.2.1 Robots as Tools

The simplest way that robots can be used is as tools to aid in the completion of physical tasks. Although there are many examples of robots used in this manner, a few examples are given that benefit from human-robot interaction. For example, agricultural robots raise the quality of produce, lower the production costs and reduce the amount on manual labor needed (Edan, 1999).

Bechar and Edan (2003) implemented a human-robot collaborative system to increase the success rate of harvesting. Results show that varying the level of autonomy resulted in improved harvesting of melons. The detection process was varied from the human independently identifying targets, to the human working with automation recognition software, and ultimately the automation software independently identifying target melons. Depending on the complexity of the harvesting environment, varying the level of autonomy of the robotic

harvester increased positive detection rates of melons by 4.5% - 7% from the human operator working alone and as much as 20% compared to autonomous robot detection alone.

Robots are often used for hazardous tasks. For instance, the placement of radioactive waste in centralized intermediate storage is best completed by robots instead of humans (Tsoukalas and Bargiotas, 1996). Robotic completion of this task in a totally autonomous fashion is desirable but not yet obtainable due to the dynamic operating conditions. Radiation surveys are completed initially through teleoperation, the learned task is then put into the robots repertoire so the next time the task is to be completed the robot will not need instruction.

A dynamic control scheme is needed so that the operator can observe the robot as it completes its task and when the robot needs help the operator can intervene and assist with execution. In a similar manner, Ishikawa and Suzuki (1997) developed a system to patrol a nuclear power plant. Under normal operation the robot is able to work autonomously. However, in abnormal situations the human must intervene to make decisions on the robots behalf. In this manner, the system has the ability to utilize the problem solving skills of the human, and thus be able to cope with unexpected events.

Human-robot teams are also used in Urban Search and Rescue (USAR). Robots are teleoperated and used mainly as tools to search for survivors. Studies completed on human-robot interaction for USAR reveal that the lack of situation awareness has a negative effect on performance (Burke et al., 2004; Murphy, 2004; Yanco and Drury, 2004; Yanco et al., 2004). Situation awareness was defined by Endsley (1988) as the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future. Burke et al. (2004) found that operators were engaged in the search task 32% of the time and the time the robot was stationary was close to 50%. These results suggest that it is extremely difficult for operators to establish situation awareness, and hence had a negative impact on performance.

The use of an overhead camera and automatic mapping techniques improve situation awareness and reduce the number of navigational errors (Scholtz, 2002; Scholtz et al., 2005). USAR is conducted in uncontrolled, hazardous

environments with adverse ambient conditions that affect the quality of sensor and video data. Studies show that varying the level of robot autonomy and combining data from multiple sensors, thus using the best sensors for the given situation, increases the success rate of identifying survivors (Nourbakhsh et al., 2005).

Ohba et al. (1999) developed a system where multiple operators in different locations control the collision free coordination of multiple robots in a common work environment. Due to teleoperation time delay and the operators being unaware of each other's intentions, a predictive graphics display was utilized to avoid collisions. The predictive display enlarged the virtual thickness of the robotic arm being controlled by other operators as a buffer to prevent collisions caused by time delay and the remote operators not being aware of each other's intentions.

In further work, operator's commands were sent simultaneously to the robot and the graphics predictor to circumvent the time delay (Chong et al., 2001). The predictive simulator used these commands to provide virtual force feedback to the operators to avoid collisions that might otherwise have occurred had the time delay not been addressed. The predictive graphics display is an important means of communicating intentions and increasing situation awareness, thus reducing the number of collisions and damage to the system.

This section on Robots as Tools highlighted two important requirements for an effective human-robot collaboration system. First, adjustable autonomy, or enabling the system to vary the level of robotic system autonomy, increases productivity and is an essential component of an effective collaboration system. Second, situation awareness, or knowing what is happening in the robot's workspace, is also essential in a collaborative system. The human member of the team must know what is happening in the robot's workspace to avoid collisions or damage to the robotic system.

2.2.2 Guide, Hosting and Assistant Robots

Robots are also used as guides. For example, Nourbakhsh et al. (1999) created and installed Sage, an autonomous mobile robot in the Dinosaur Hall at the Carnegie Museum of Natural History. Sage interacts with museum visi-

tors through an LCD screen and audio, and uses humor to creatively engage visitors. Sage also exhibits emotions and changes in mood to enhance communication. Sage is completely autonomous and when confronted with trouble will stop and ask for help.

Sage was designed with safety, reliability and social capabilities to enable it to be an effective member of the museum staff. Sage shows not only how speech capabilities affect communication, but also that the form of speech and non-verbal communication influences how well communication takes place. These capabilities are demonstrated by Sage reacting in a different manner to the same situation. For example, if someone repeatedly stands in its way, Sage will raise its voice if it is "annoyed" or, if Sage is "happy", it will respond by joking around and attempting to engage the person. Sage is shown in Figure 2.1.

Figure 2.1 Sage interacting through an LCD screen with museum visitors (Nourbakhsh et al., 1999).

The autonomous interactive robot Robovie is a humanoid robot that communicates and interacts with humans as a partner and guide (Kanda et al., 2002). Its use of gestures, speech and eye contact enables the robot to effec-

tively communicate with humans. Results of experiments showed that robot communication behavior induced human communication responses that increased understanding. During interaction with Robovie participants spent more than half of the time focusing on the face of the robot indicating the importance of gaze in human-robot communication. Robovie is shown in Figure 2.2.

Figure 2.2 Robovie interacting with school children (Kanda et al., 2002).

Robots used as guides in museums must interact with people and portray human-like behavior to be accepted. Kuzuoka et al. (2004) conducted studies in a science museum to see how humans predict or anticipate the unfolding of events when they communicate. The term projection was used to describe the capacity to predict or anticipate the unfolding of events. Projection was found to be difficult through speech alone because speech does not allow a partner to anticipate what the next action may be in the way a person can predict what may happen next by body language (gesture) or focus point of gaze.

Kuzuoka et al. (2004) designed a remote instruction robot, Gestureman, to investigate projectability properties. A remote operator, who was located in a separate room, controlled Gestureman. Through Gestureman's three cameras the remote operator had a wider view of the local work space than a person normally would and so could see objects without the robot facing them, as

shown in Figure 2.3. This dual ecology led to local human participants being misled as to what the robot was focusing on, and thus not being able to quickly locate what the remote user was trying to identify. The experiment highlighted the importance of gaze direction and situation awareness in effective remote collaboration and communication.

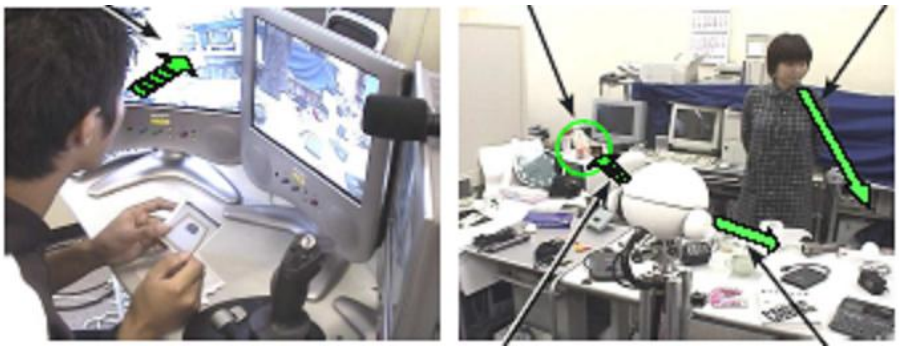


Figure 2.3 Gestureman: Remote user (left) with wider field of view than robot, identifies object but does not project this intention to local participant (right) (Kuzuoka et al., 2004).

Sidner and Lee (2005) show that a hosting robot must not only exhibit conversational gestures, but also must interpret these behaviors from their human partner to engage in collaborative communication. Their robot Mel, a penguin hosting robot, uses vision and speech recognition to engage a human partner in a simple demonstration. Mel points to objects in the real world, tracks the gaze direction of the participant to ensure instructions are being followed, and looks at observers to acknowledge their presence.

Mel actively participates in the conversation and disengages from the conversation when appropriate. Mel is a good example of a multimodal approach to effectively ground a conversation. More explicitly, gesture, gaze direction and speech are used to ensure two-way communication is taking place. Mel is shown in Figure 2.4.

Cero (Huettenrauch et al., 2004) is an assistant robot designed to help those with physical disabilities in an office environment. During the iterative development of Cero, user studies showed that communicating through speech alone was not effective enough. Users commented that they could not distinguish where the front of the robot was nor could they determine if their commands to the robot were understood correctly. In essence, communication was not being effectively grounded.

Figure 2.4 Mel uses multimodal communication to interact with participants (Sidner and Lee, 2005).

To overcome this difficulty, a humanoid figure was mounted on the front of the robot that could move its head and arms, as shown in Figure 2.5. After implementation of the humanoid figure, it was found that users felt more comfortable communicating with the robot and grounding was easier to achieve (Huettenrauch et al., 2004). These results highlight the importance of grounding in communication and the impact that gestures have on grounding.

Figure 2.5 Cero robot with humanoid figure using gestures to enhance grounding (Huettenrauch et al., 2004).

An assistant robot should exhibit a high degree of autonomy to obtain information about their human partner and surroundings. Iossifidis et al. (2003) developed CoRa (Cooperative Robot Assistant) that is modeled on the behaviors, senses, and anatomy of humans. CoRa is fixed on a table and interacts through speech, hand gestures, gaze and mechanical interaction allowing it to obtain the necessary information about its surrounding and partner. CoRa's tasks include visual identification of objects presented by its human teacher, recognition of an object amongst many, grasping and handing over of objects and performing simple assembly tasks.

Lessons learned from this section for the design of an effective human-robot collaboration system include the need for effective natural speech. Therefore, a multimodal approach is necessary as communication is more than just speech alone. In addition, the communication behaviour of a robotic system is important as it should induce natural communication with human team members. Lastly, grounding is a key element in communication, and thus collaboration.

2.2.3 Humanoid Robots

Robonaut is a humanoid robot designed by NASA to be an assistant to astronauts during an extra vehicular activity (EVA) mission. Interaction with Robonaut occurs in the three roles outlined in the work on human-robot interaction by Scholtz (2003): 1) remote human operator, 2) a monitor and 3) a coworker. Robonaut is shown in Figure 2.6.

Robonaut's anthropomorphic form allows an intuitive one to one mapping for remote teleoperation. To enhance the operator's sense of immersion feedback is provided in the form of visual aids and kinesthetic, tactile and auditory cues (Glassmire et al., 2004). The co-worker interacts with Robonaut in a direct physical manner, as shown in Figure 2.7, and therefore is much like interacting with a human.

Research into humanoid robots has also concentrated on making robots appear human in their behavior and communication abilities. For example, Breazeal et al. (2001) are working with Kismet, a robot that has been endowed with visual perception that is human-like in its physical implementation. Eye movement and gaze direction play an important role in communication aiding

Figure 2.6 Robonaut with coworker and remote human operator (Glassmire et al., 2004).

Figure 2.7 Robonaut interacting in a direct manner with coworker (Glassmire et al., 2004).

the participants in reaching common ground. By following the example of human vision movement and meaning, Kismet's behavior will be understood and Kismet will be more easily accepted socially. Kismet is an example of a robot that can show the non-verbal cues typically present in human-human conversation. Kismet is shown in Figure 2.8.

Figure 2.8 Kismet displaying non-verbal communication cues (Breazeal et al., 2001).

Robots with human social abilities, rich social interaction and natural communication will be able to learn from human counterparts through cooperation and tutelage. Breazeal et al. (Breazeal, 2004; Breazeal et al., 2003) are working toward building socially intelligent cooperative humanoid robots that can work and learn in partnership with people. Robots will need to understand intentions, beliefs, desires and goals of humans to provide relevant assistance and collaboration. To collaborate, robots will also need to be able to infer and reason.

The goal is to have robots learn as quickly and easily, as well as in the same manner, as a person. Their robot, Leonardo, is a humanoid designed to express and gesture to people, as well as learn to physically manipulate objects from natural human instruction, as shown in Figure 2.9. The approach for Leonardo's learning is to communicate both verbally and non-verbally, use visual deictic references, and express sharing and understanding of ideas with its teacher.

Figure 2.9 Leonardo activating middle button (left) and learning the name of the left button (right) (Breazeal et al., 2001).

2.2.4 Robots in Collaborative Tasks

Inagaki et al. (1995) have proposed that humans and robots can have a common goal and work cooperatively through perception, recognition and intention inference. One partner would be able to infer the intentions of the other from language and behavior during collaborative work. Morita et al. (1998) demonstrated that the communication ability of a robot improves with physical and informational interaction synchronized with dialog. Their robot, Hadaly-2, expresses efficient physical and informational interaction and is capable of carrying an object to a target position by reacting to visual and audio instruction.

Natural human-robot collaboration requires the robotic system to understand spatial referencing. Tversky et al. (1999) observed that in human-human communication, speakers used the listener's perspective when the listener had a higher cognitive load than the speaker. Tenbrink et al. (2002) presented a method to analyze spatial human-robot interaction, in which natural language instructions were given to a robot via keyboard entry. Results showed that the humans used the robot's perspective for spatial referencing.

To allow a robot to understand different reference systems, Roy et al. (2004) created a system where their robot was capable of interpreting the environment from its perspective or from the perspective of its conversational partner. Using verbal communication, their robot Ripley was able to under-

stand the difference between spatial references such as “my left” and “your left”. The results of Tenbrink et al. (2002), Tversky et al. (1999) and Roy et al. (2004) illustrate the importance of situation awareness and a common frame of reference in spatial communication.

Skubic et al. (2004, 2002) also conducted a study on human-robotic spatial dialog. A multimodal interface was used, including speech, gestures, sensors and personal electronic devices. The robot was able to use dynamic levels of autonomy to reassess its spatial situation in the environment through the use of sensor readings and an evidence grid map. The result was natural human-robot spatial dialog enabling the robot to communicate obstacle locations relative to itself and receive verbal commands to move to an object it had detected.

Rani et al. (2004) built a robot that senses the anxiety level of a human using biofeedback sensors and responds appropriately. In dangerous situations, where the robot and human are working in collaboration, the robot was able to detect the anxiety level of the human and take appropriate actions. To minimize bias or error, the emotional state of the human is interpreted by the robot through physiological responses that are generally involuntary and are not dependent upon culture, gender or age.

To obtain natural human-robot collaboration, Horiguchi et al. (2000) developed a teleoperation system where a human operator and an autonomous robot share their intent through a force feedback system. The human or the robot can control the system while maintaining their independence by relaying their intent through a force feedback joystick. Both the human and robot affect the feedback on the joystick whose position is what ultimately drives the robot. The use of force feedback resulted in reduced execution time and fewer stalls of the teleoperated mobile robot.

Fernandez et al. (2001) also introduced an intention recognition system where a robot participating in the transportation of a rigid object detects a force signal measured in the arm gripper. The robot uses this force information, as non-verbal communication, to generate its motion planning to collaborate in the execution of the transportation task. Force feedback used for intention recognition is another way in which humans and robots can communicate non-verbally and work together.

Collaborative control was developed by Fong et al. (2002a,b, 2003) for mobile autonomous robots. The robots work autonomously until they run into a problem they can't solve. At this point, the robots ask the remote operator for assistance, allowing human-robot interaction and autonomy to vary as needed. Performance deteriorates as the number of robots working in collaboration with a single operator increases (Fong et al., 2003). Conversely, robot performance increases with the addition of human skills, perception and cognition, and benefit from human advice and expertise.

In the collaborative control structure used by Fong et al. (2002a,b, 2003), the human and robots engage in dialog, exchange information, ask questions and resolve differences. Thus, the robot has more freedom in execution and is more likely to find good solutions when it encounters problems. More succinctly, the human is a partner whom the robot can ask questions, obtain assistance from and in essence, collaborate with.

In more recent work, Fong et al (Fong et al., 2006) note that for humans and robots to work together as peers, the system must provide mechanisms for the humans and robots to communicate effectively. They introduced the Human-Robot Interaction Operating System (HRI/OS) which enables a team of humans and robots to work together on tasks that are well defined and narrow in scope. The human agents are able to use spatial dialog to communicate and the autonomous agents use spatial reasoning to interpret "left of" type elements from the spatial dialog. The ambiguities arising from such dialog are resolved through the use of modeling the situation in a simulator.

Research has shown that for robots to be effective partners they should interact meaningfully through mutual understanding. In addition, a human-robot collaborative system should take advantage of varying levels of autonomy and multimodal communication allowing the robotic system to work independently and ask its human counterpart for assistance when a problem is encountered. Communication cues should be used to help identify the focus of attention, greatly improving performance in collaborative work. Finally, grounding can be achieved through meaningful interaction and the exchange of dialog.

2.3 Summary

The review of research in communication and HRI has resulted in a number of requirements for creating an effective human-robot collaboration system:

- Grounding is a key element in communication, and thus in collaboration.
- Use of effective natural speech and a multimodal approach is necessary as communication is more than just speech alone.
- The communication behaviour of a robotic system is important, as it should induce natural communication with human team members.
- To be an effective partner, a robot should interact meaningfully through mutual understanding.
- Communication cues should be used to help identify the focus of attention, greatly improving performance in collaborative work.
- Adjustable autonomy increases productivity and is an essential component of an effective collaboration system.
- The robotic system should work independently and ask its human counterpart for assistance when a problem is encountered.
- Maintaining situation awareness is essential in any collaboration system. The human member of the team must know what is happening in the robot's world to avoid collisions or damage to the robotic system.

Chapter 3

Augmented Reality for Human-Robot Collaboration

Augmented Reality (AR) is used in this research to enable humans to effectively communicate with a robotic system by providing a platform for easily reaching common ground. Through the use of AR, an environment is created that is rich with spatial cues and is thus potentially more conducive to collaboration.

This chapter begins by introducing AR and presents work describing the use of AR in human-human collaborative tasks. It then examines the use of AR in current HRI research. A summary is provided listing the benefits of AR for this task and how an effective human-robot collaboration system can be created by taking advantage of these benefits.

A review of multimodal interaction is then provided and a discussion is given of the Wizard of OZ study technique which was employed in this research. The chapter finishes by summarizing the lessons learned from this chapter and the previous chapter on HRI. This summary highlights the components necessary for an effective human-robot collaboration system. A second focus is on how the benefits of using AR technology can help to make this type of human-robot collaborative system a reality.

3.1 Augmented Reality

3.1.1 Introduction to Augmented Reality

Augmented Reality (AR) is a technology that facilitates the overlay of computer graphics onto the real world view of the user. AR differs from virtual reality (VR) in that it uses graphics to augment the physical world rather than replacing it entirely, as in a virtual environment. Therefore, AR enhances rather replaces reality. Azuma et al. (2001) note that AR computer interfaces have three key characteristics:

- They combine real and virtual objects.
- The virtual objects appear registered on the real world.
- The virtual objects can be interacted with in real time.

In a typical AR system, the user wears a head mounted display (HMD) with a camera mounted on it. The output from the camera is fed into a computer, augmented with 3D graphics and then fed back into the HMD. Therefore, the user sees an enhanced view of the real world through the video image in the HMD. This type of AR set-up is commonly called a video-see-through AR interface and is shown in Figure 3.1.

To precisely overlay virtual images onto the real world view, it is necessary to track the user's viewpoint. One way of doing this is through the use of computer vision. Square fiducial patterns are placed in the real environment with a unique symbol in the middle of each pattern. Computer vision techniques are then used to identify the unique symbols, calculate the camera position and orientation from these symbols, and display 3D virtual images aligned with the position of the fiducial patterns. This technique creates an augmented view of the real world and is made possible through the use of the ARToolKit computer vision tracking library (ARToolKit, 2008). The ARToolKit library calculates camera position at more than 30 frames per second and to millimeter level accuracy.

The augmented view is then fed into the HMD providing the user with a seamless combination of the real world view and virtual graphics, where the

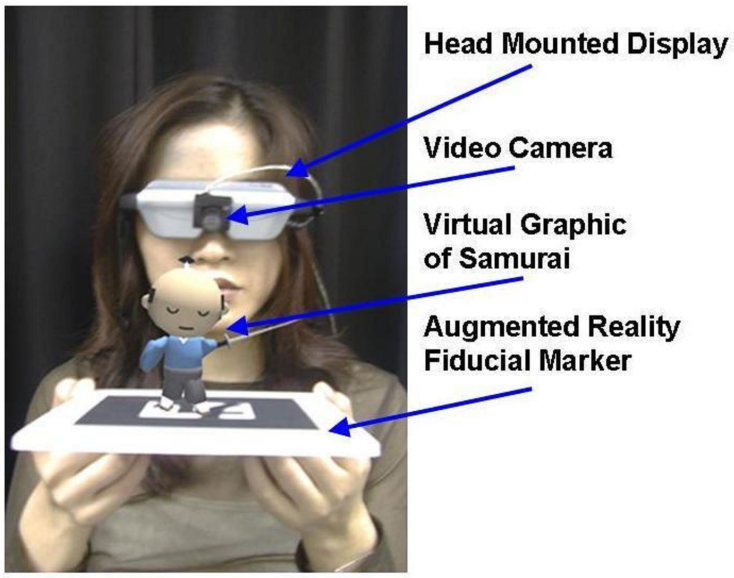
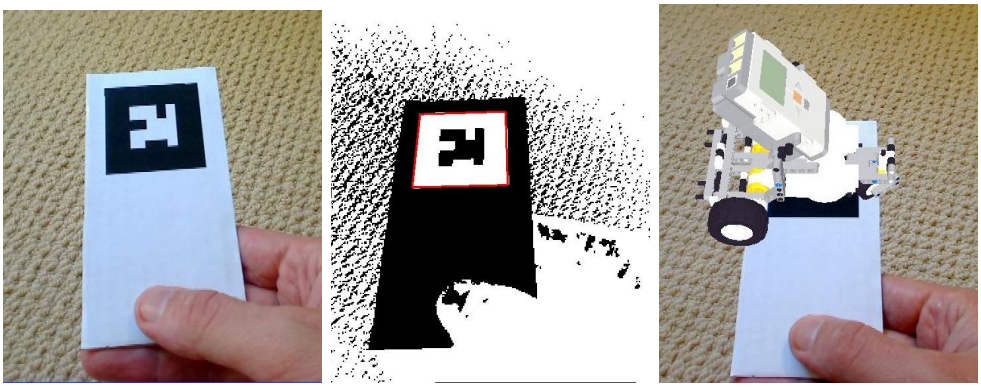


Figure 3.1 Video see-through AR interface (Billinghurst et al., 2000).

virtual images appear fixed to the fiducial patterns. This process is depicted in Figure 3.2. Therefore, AR blends virtual 3D graphics with the real world in real time (Azuma, 1997). In addition, the ability to manipulate the physical markers with fiducial patterns on them enables direct real-time interaction with the 3D virtual content (Billinghurst et al., 2005). AR also supports transitional user interfaces along the entire spectrum of Milgram's Reality-Virtuality continuum (Milgram and Kishino, 1994), see Figure 3.3. Therefore, AR enables a smooth transition from reality to virtuality.

One way to support interaction with AR content is through the use of a Tangible User Interface metaphor. Tangible User Interfaces (TUIs) use real-world objects as the interaction devices for a computer (Ishii and Ullmer, 1997). Therefore, TUIs are extremely intuitive to use because physical object manipulations are mapped one-to-one to virtual object operations (Fitzmaurice and Buxton, 1997). Another benefit of a TUI is that it naturally supports sharing and collaboration.

TUIs are a viable approach for interaction with AR applications as they enable users to interact naturally by manipulating real world objects. Thus, the principles of TUIs can be combined with AR's display capabilities in an interface metaphor known as Tangible Augmented Reality (TAR) (Kato et al.,



(a) Video capture

(b) Fiducial marker identification, position and orientation calculation

(c) Virtual image aligned with marker, presented to user in HMD

Figure 3.2 AR Video See Through Process.

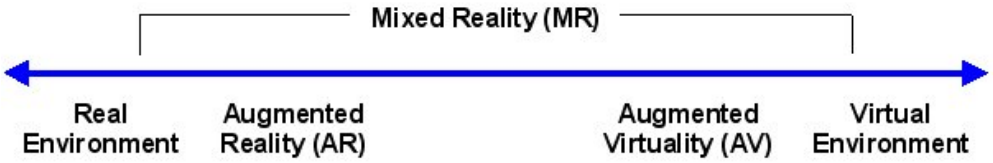


Figure 3.3 Milgram's Reality Virtuality Continuum (Milgram and Kishino, 1994).

2001). A TAR interface supports the presentation of 3D virtual objects anywhere in the physical environment, while simultaneously allowing users to interact with this virtual content using real world physical objects (Kato et al., 2000). An ideal TAR interface facilitates seamless display and interaction, removing the functional and cognitive seams found in traditional AR and TUI interfaces.

3.1.2 AR in Collaborative Tasks

AR technology can be used to enhance face-to-face collaboration. For example, the Shared Space application effectively combined AR with physical and spatial user interfaces in a face-to-face collaborative environment (Billinghurst et al., 2000). In this game, manipulation of physical markers with square fiducial patterns on them was used for interaction with virtual content. Through the ability of the ARToolKit (2008) software to track the physical markers, users

were able to interact with and exchange markers, thus effectively collaborating in a 3D AR environment. When two corresponding markers were brought together, it would result in an animation being played. For example, when a marker with an AR depiction of a witch was put together with a marker with a broom, the witch would jump on the broom and fly around.

User studies found that people had no difficulties using the system to interact together, displaying collaborative behavior seen in typical face-to-face interactions (Billinghurst et al., 2000). The Shared Space application supports natural face-to-face communication by allowing multiple users to see each other's facial expressions, gestures and body language, demonstrating that a 3D collaborative environment enhanced with AR content can seamlessly enhance face-to-face communication and allow users to naturally work together, as shown in Figure 3.4.

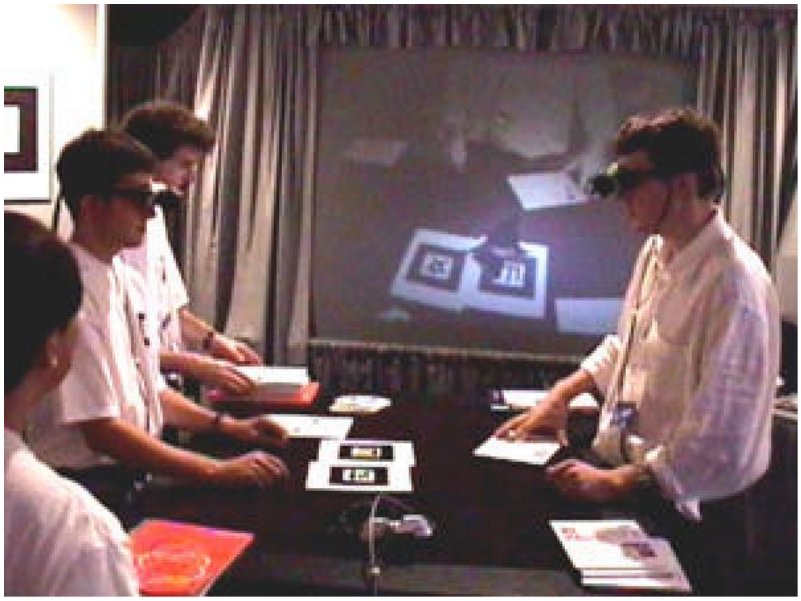


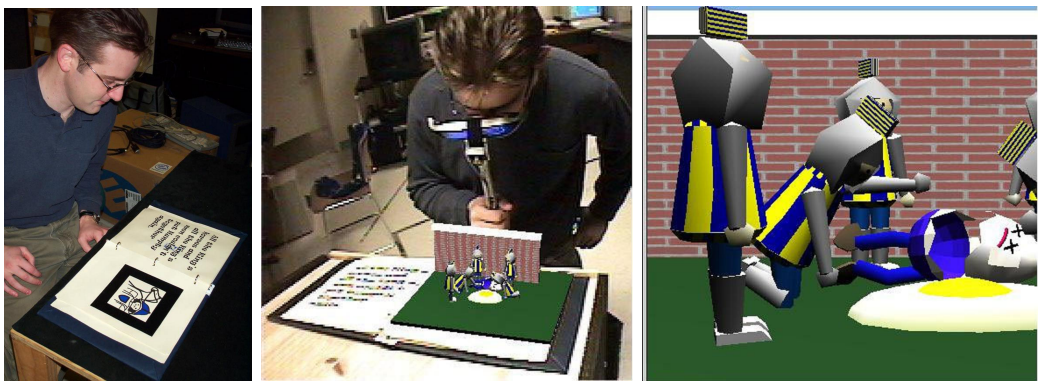
Figure 3.4 The Shared Space application enables users to see and interact with physical and virtual objects and see other participants, creating an effective collaborative environment (Billinghurst et al., 2000).

Another example of the ability of AR to enhance collaboration is the MagicBook application. The MagicBook allows for a continuous seamless transition from the physical world to augmented and/or virtual reality (Billinghurst et al., 2001). It utilizes a real book that can be read normally, or one can use a hand held display (HHD) to view AR content popping out of the real book.

The placement of the augmented scene is achieved by the ARToolKit (2008) computer vision library.

When a user is interested in a particular AR scene they can fly into the scene and experience it as an immersive virtual environment by simply flicking a switch on the hand held display. Once immersed in the virtual scene, when a user turns their body in the real world, the virtual viewpoint changes accordingly. The user can also fly around in the virtual scene by pushing a pressure pad in the direction they wish to fly. When the user switches to the immersed virtual world an inertial tracker is used to place the virtual objects in the correct location.

The MagicBook application also supports multiple simultaneous users who each see the virtual content from their own viewpoint. When the users are immersed in the virtual environment they can experience the scene from either an ego-centric or exo-centric point of view (Billinghurst et al., 2001). The MagicBook thus provides an effective environment for collaboration by allowing users to see each other when viewing the AR application, maintaining important visual cues needed for effective collaboration. When immersed in the VR environment, users are represented as virtual avatars and can be seen by other users in the AR or VR scene, thereby maintaining awareness of all users, and thus still providing an environment supportive of effective collaboration. The MagicBook application is shown in Figure 3.5.



(a) Reality (b) Augmented Reality (c) Virtual Reality

Figure 3.5 Using the MagicBook to move from Reality to Virtual Reality (Billinghurst et al., 2001).

Prince et al. (2002a,b) introduced a 3D live augmented reality conferencing system. Through the use of multiple cameras and an algorithm determining shape from silhouette, they were able to superimpose a live 3D image of a remote collaborator onto a fiducial marker, creating the sense that the live remote collaborator was in the workspace of the local user. The shape from silhouette algorithm works by each of 15 cameras identifying a pixel as belonging to the foreground or background, and then isolation of the foreground information produces a 3D image that can be viewed from any angle by the local user. Figure 3.6 shows the live collaborator displayed on a fiducial marker.

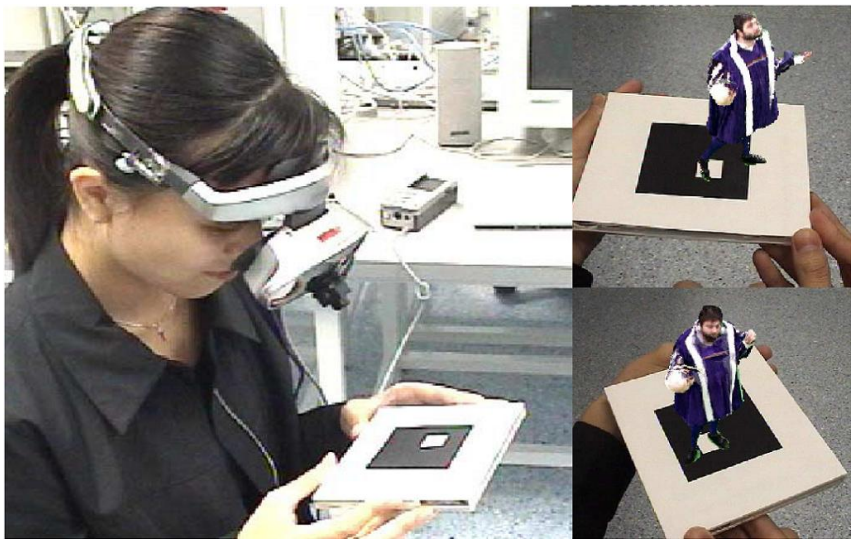


Figure 3.6 Remote collaborator as seen on AR fiducial marker (Prince et al., 2002b).

Cheok et al. (2002) utilized shape from silhouette live 3D imagery (Prince et al., 2002b) and wearable computers to create an interactive theater experience. Their outdoor mobile AR setup is shown in Figure 3.7 and the interactive theater experience is shown in Figure 3.8. Participants collaborate in both an indoor and outdoor setting. Users seamlessly transition between the real world, augmented and virtual reality, allowing multiple users to collaborate and experience the theater interactively with each other and 3D images of live actors.

The Human Pacman game (Cheok et al., 2003) is an outdoor mobile AR application that supports collaboration. The system allows for mobile AR users to play together, as well as get help from stationary observers. Human Pacman supports the use of tangible and virtual objects as interfaces for the

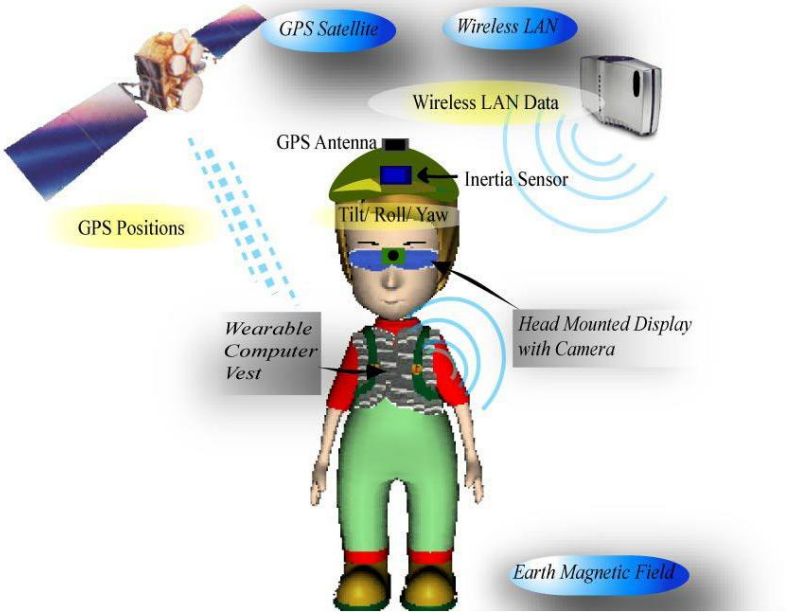


Figure 3.7 Mobile AR Setup (Cheok et al., 2002).

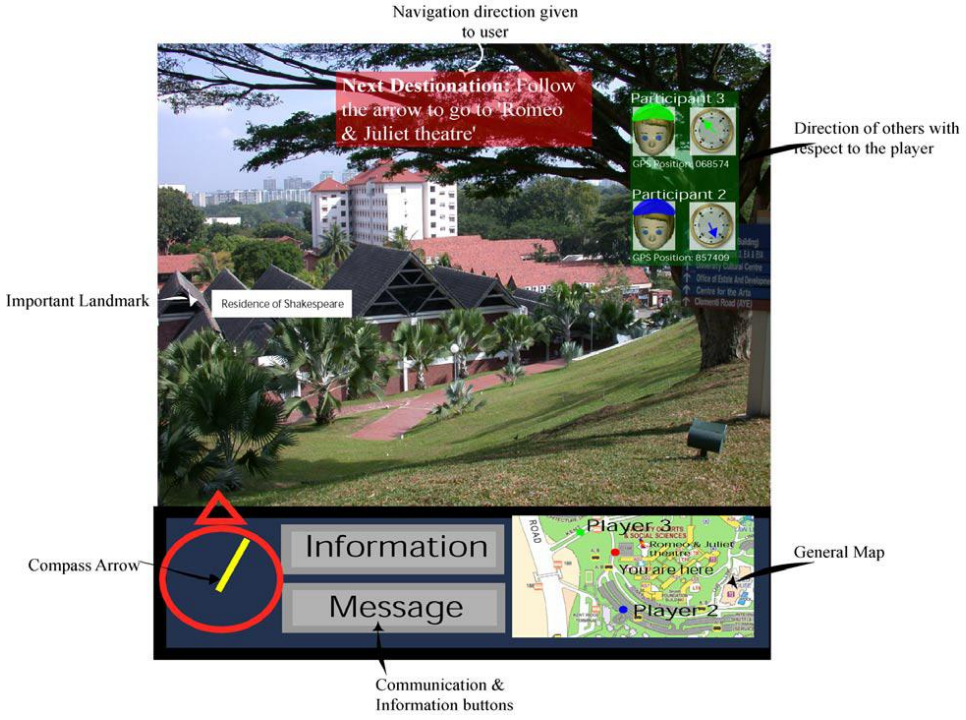


Figure 3.8 Interactive Theater Experience (Cheok et al., 2002).

AR game, as well as allowing real world physical interaction between players. Players are able to seamlessly transition between a first person augmented reality world and an immersive virtual world. The use of AR allows the virtual Pacman world to be superimposed over the real world setting. AR enhances collaboration between players by allowing them to exchange virtual content as they are moving through the AR outdoor world. The Human Pacman application can be seen in Figure 3.9.

Figure 3.9 AR Human Pacman game (Cheok et al., 2003).

Reitmayr and Schmalstieg (2004) implemented a mobile AR tour guide system that allows multiple tourists to collaborate while they explore a part of the city of Vienna, the system in shown in Figure 3.10. Their system directs the user to a target location and displays location specific information that can be selected to provide detailed information. When a desired location is selected, the system computes the shortest path, and displays this path to the user as cylinders connected by arrows, as shown in Figure 3.11. Their system helps the user to maintain situation awareness in unfamiliar surroundings.

Kiyokawa et al. (2002) experimented with how diminished visual cues of co-located users in an AR collaborative task influenced task performance. Performance was best when collaborative partners were able to see each other in

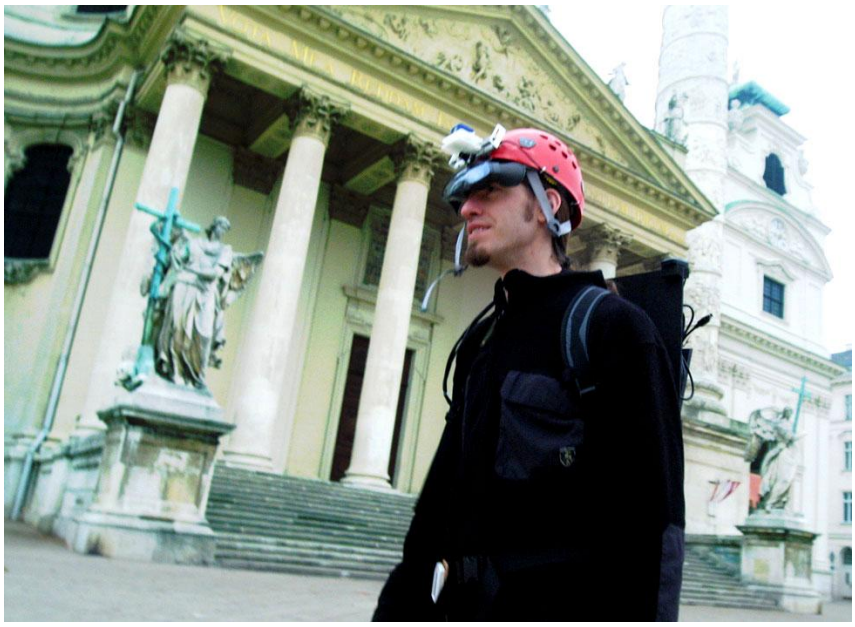


Figure 3.10 Mobile AR Tour Guide System (Reitmayr and Schmalstieg, 2004).

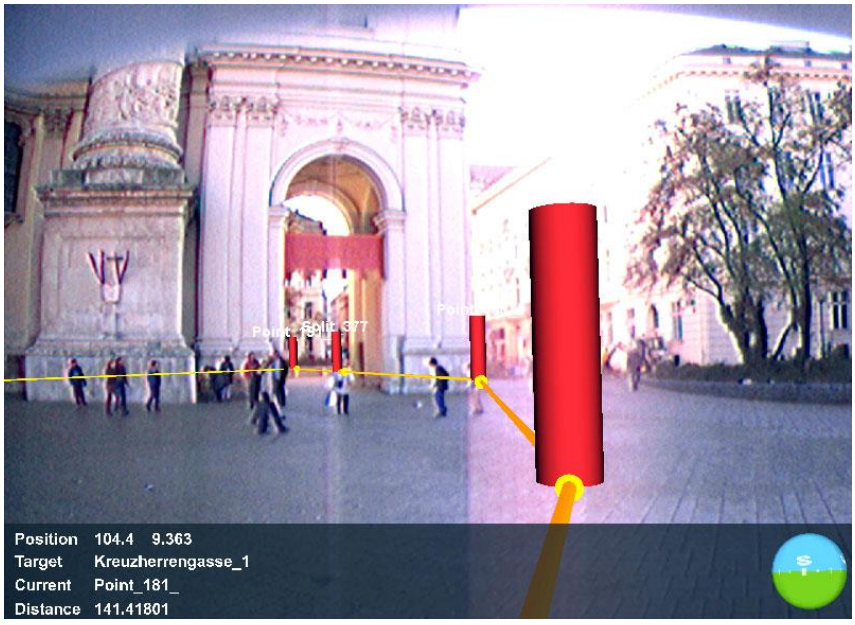


Figure 3.11 AR Tour Guide displaying path to follow (Reitmayr and Schmalstieg, 2004).

real time. The worst case occurred in an immersive virtual reality environment where the participants could only see virtual images of their partners.

In a second experiment, Kiyokawa et al. (2002) modified the location of the task space, as shown in Figure 3.12. Participants expressed more natural communication when the task space was between them. However, the orientation of the task space was significant. The task space between the participants meant that one person had a reversed view from the other. Results showed that participants preferred the task space to be on a wall to one side of them, where they could both view the workspace from the same perspective. The results of this research highlight the importance of the task space location, the need for a common reference frame and the ability to see the visual cues displayed by a collaborative partner.

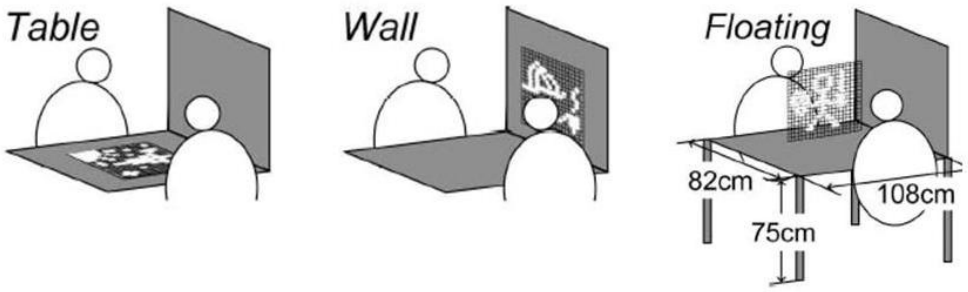


Figure 3.12 Different locations of task space in Kiyokawa et al second experiment (Kiyokawa et al., 2002).

The results from this section show that AR can enhance face-to-face collaboration through allowing the use of physical tangible objects for ubiquitous computer interaction. Therefore, making the collaboration natural by allowing participants to use objects for interaction that they would normally use in a collaborative effort. Additionally, AR provides rich spatial cues permitting users to interact freely in space, supporting the use of natural spatial dialog.

Collaboration is also enhanced by the use of AR since facial expressions, gestures and body language are effectively transmitted. Multiple users can view the same virtual content from their own perspective, either from an ego- or exo-centric viewpoint. AR also allows users to see each other while viewing the virtual content, thereby maintaining spatial awareness. The position of the workspace in an AR environment can be optimized to enhance collaboration.

3.1.3 AR in Human-Robot Interaction

There has been some previous work on using AR to enhance HRI. For example, Milgram et al. (1993) pointed out the need for HRI systems that can transfer the interaction mechanisms that are considered natural for human communication to the precision required for machine information. Their approach was to use augmented overlays in a fixed work environment. These graphic overlays enabled the human "director" to use spatial referencing to interactively plan and optimize the path of a robotic manipulator arm, see Figure 3.13.

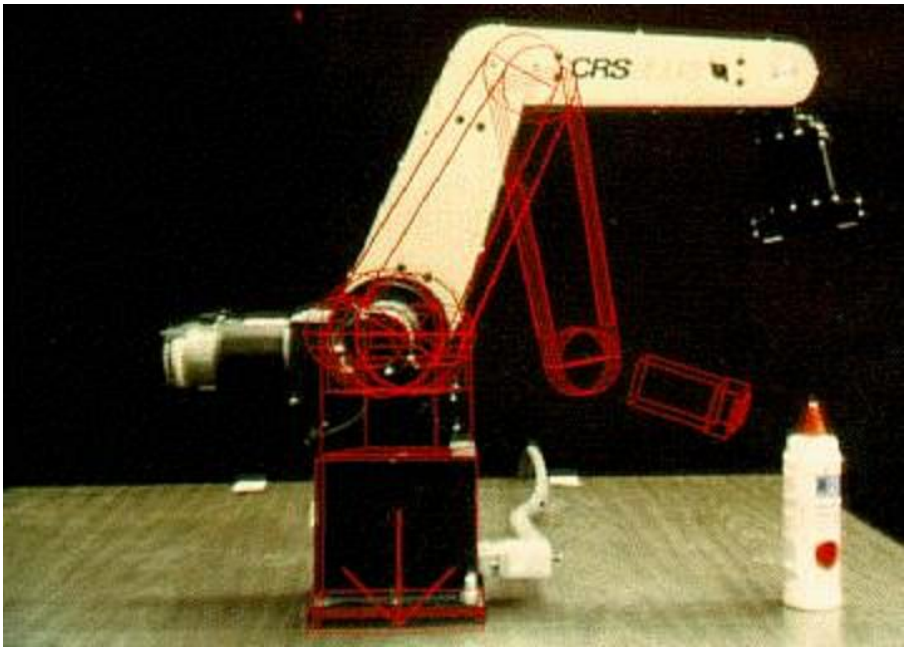


Figure 3.13 AR overlay in fixed environment for interactive path planning (Milgram et al., 1993).

Milgram et al. (1993) also highlighted the need for combining the attributes that humans are good at with those that robots are good at to create an optimized human-robot team. For example, humans are good at approximate spatial referencing, such as using the ambiguous terms "here" and "there" while pointing to a location in 3D space. However, robotic systems need highly accurate discrete information.

Giesler et al. (2004) implemented an AR system that creates a path for a mobile robot to follow using voice commands and a wand. The wand had fiducial markers attached to it so that the ARToolKit (2008) could be used

for tracking and interaction in the AR environment. Fiducial markers were placed on the floor and used to calibrate the tracking coordinate system. A path was created node by node, by pointing the wand at the floor and giving voice commands for the meaning of a particular node. Map nodes could be interactively moved or deleted.

The robot moved from node to node using its autonomous collision detection capabilities. As goal nodes were reached, the node depicted in the AR system changed colour to keep the user informed of the robots progress. The robot would retrace steps if an obstruction was encountered and would then create a new plan to arrive at the goal destination, as shown in Figure 3.14. Although Giesler et al. (2004) did not mention a user evaluation, they did comment that the interface was intuitive to use. Results from their work show that AR is an excellent application to visualize planned trajectories and inform the user of the robots progress and intention.

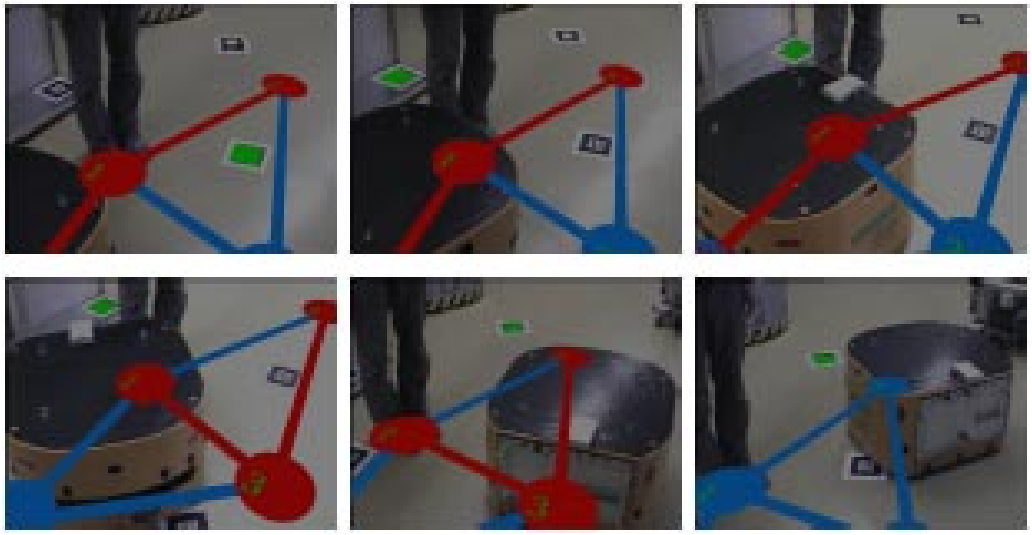


Figure 3.14 Robot follows AR path nodes, redirects when obstacle in the way (Giesler et al., 2004).

Maida et al. (2007) conducted a study of an alignment task using a robotic manipulator arm. One condition made use of AR overlay information to help the user guide the arm and the other did not. Results from the study showed a significant improvement in performance for the condition with AR overlays. Similarly, Drury et al. (2006) found that for operators of Unmanned Aerial Vehicles (UAVs) augmenting real-time video with preloaded map terrain data

made it significantly easier to understand 3D spatial relationships compared to using 2D video alone. The augmented video resulted in increased situation awareness of the activities of the UAV. AR has also been used to display robot sensor information on the view of the real world (Collett and MacDonald, 2006).

The results from this section show that AR can be used to transfer the approximate spatial referencing natural in human communication to the discrete positional information required by robotic systems. AR has also been shown to be effective in visualizing robot plans and informing the user of the robot's progress in completing a plan. The use of AR has also been shown to increase performance in robotic control and is capable of increasing situation awareness. Finally, through the augmented overlays of the real world environment, it is possible to keep the user informed of the internal state of the robotic system.

3.1.4 Summary

Augmented Reality is an ideal platform for human-robot collaboration as it provides many of the features required for robust communication and collaboration. Benefits of the use of AR for this type of application include:

- The ability for a human to share a remote (ego-centric) view with a robot, thus enabling the ability for the human-robot team to reach common ground.
- The ability for the human to have a world view (exo-centric) of the collaborative workspace, thus affording spatial awareness.
- The use of deictic gestures and spatial dialog by allowing all partners to refer to and interact with the graphic 3D overlaid imagery, supporting the use of natural spatial dialog.
- Collaboration of multiple users, multiple humans can effectively collaborate with multiple robotic systems.
- Seamless transition from the real world to an immersive data space that aids in the grounding process and increases situation awareness.

- Display of visual cues as to what the robots intentions are and it's internal state, greatly enhancing the grounding process and increasing situation awareness.

- Providing the spatial cues necessary for both local and remote collaboration.

Overall, a human-robot collaboration system would benefit greatly from the use of AR. AR could enhance the grounding process, provide for increased situation awareness, enable the use of natural spatial dialog, allow for multiple collaborative partners and enable both local and remote collaboration. The result would be a system that allows natural and effective communication and thus collaboration.

More specifically, multiple users can view the same fiducial patterns and therefore have their own perspective of the 3D virtual content. Since the users see each other's facial expressions, gestures and body language, AR therefore supports natural face-to-face communication. This interaction demonstrates that a 3D collaborative environment enhanced with AR content can seamlessly enhance face-to-face communication and allow users to naturally work together (Billinghurst et al., 2001, 2000). Shared visual workspaces of this type have been shown to enhance collaboration, as they increase situation awareness (Fussell et al., 2003).

AR can provide a virtual 3D world model that both the human and robotic system can operate within. This use of a common 3D world enables both the human and robotic system to utilize the same common reference frames. The use of AR will support the use of spatial dialog and deictic gestures, allows for adjustable autonomy by supporting multiple human users, and will allow the robot to visually communicate to its human collaborators its internal state through graphic overlays on the real world view of the human. The use of AR enables a user to experience a tangible user interface, where physical objects are manipulated to affect changes in the shared 3D scene (Billinghurst et al., 2005), thus allowing a human to reach into the 3D world of the robotic system and manipulate it in a way the robotic system can understand.

3.2 Multimodal Interaction

A multimodal system supports two or more combined user inputs, such as speech and gesture. Users have a strong preference to interact multimodally and their use of this interface style improves performance (Oviatt, 2000). Multimodal interfaces can function in a more robust and stable manner than unimodal systems that involve a single recognition technology, such as speech, pen input or vision (Oviatt, 2003). Therefore, for an HRI system to be robust and natural for the human user, a multimodal interface design is desirable.

One of the first interfaces to support multimodal speech and gesture input was the Media Room (Bolt, 1980). The Media Room allowed the user to interact with a computer through voice, gesture and gaze. Bolt's work showed that gestures combined with natural speech (multimodal interaction) lead to a powerful and more natural human-machine interface.

Work by Hauptmann (1989) investigated the use of multimodal interaction for a simple 3D cube manipulation task. The study had three conditions: participants used gestures only, speech only and speech and gestures combined. The analysis showed that people strongly preferred using a combination of speech and gestures for graphics manipulation.

Multimodal interfaces can be very intuitive because the strengths of gesture input compliment the limitations of speech input, and vice versa. Cohen (Cohen, 1992; Cohen et al., 1989) showed how speech interaction is ideally suited for descriptive techniques, while gestural interaction is ideal for direct manipulation of objects. Speech and gesture thus compliment each other and when used together create an interface more powerful than either modality alone.

Unlike gesture input, voice is not tied to a spatial metaphor (Schmandt et al., 1990) and so can be used to interact with objects regardless of whether they can be seen or not. However, care must be taken to map the appropriate modality to the application input parameters. For example, the difficulty of using speech alone was demonstrated by Kay (1993) who constructed a speech driven interface for a drawing program. Even simple cursor movements around the screen required a time consuming combination of continuous and discrete vocal commands.

A multimodal interface fusing two or more information sources can effectively reduce recognition uncertainty and stabilize system performance (Oviatt, 2003). Oviatt (2003) also showed that performance advantages were demonstrated for different modality combinations, as well as for different user groups, applications, and environments. Most importantly, the error suppression achievable with a multimodal interface, compared with a unimodal one, can be substantial.

The use of wearable computers introduces a new issue, the difficulty of using traditional desktop input devices, such as a mouse. To overcome this issue Kolsch et al. (2006) took advantage of multimodal input for their wearable AR system. They combined input from hand gestures, trackball input, voice and head pose to manipulate content in an AR setting. They conclude that multimodal user interfaces can broaden the diverse input needs of mobile applications (Kolsch et al., 2006). However, it should be noted that no formal evaluation was mentioned.

To investigate the use of a multimodal AR interface for industrial assembly, Siltanen et al. (2007) implemented a simple assembly task of putting together a 3D puzzle. Through a speech and gesture interface the user was able to receive direction on the sequence of putting the puzzle together. A limited user study (five users) showed that the multimodal interface was judged favorably. However, the users found the gestures to be exhausting as they had to raise their arms to gesture to the system. Also, the system did not provide enough feedback to the user to let them know their multimodal input was understood. Therefore, a multimodal interface can be more intuitive, but the implementation of such an interface has to be done in an effective manner, otherwise the benefits of the multimodal interface will be lost.

Some of the key lessons learned from multimodal systems include:

- Multimodal systems are more robust than unimodal systems.
- Multimodal interaction leads to a more powerful and natural human-machine interface.
- User studies have shown that participants prefer multimodal interaction.

3.3 Wizard of Oz Study

A Wizard of Oz (WOZ) study was conducted as part of this research to help define the types of speech and gestures that would be used when a human collaborates with a mobile robot. A WOZ study is one where a system in development is not yet fully functional and a human "wizard" acts for the parts of the system that have not yet been implemented. Such a study provides insight into how a system should be developed to optimize usefulness and usability. This section provides background information of what characterizes a WOZ study and its benefits.

A WOZ study is a viable means of determining how a multimodal system should function before that system is fully developed (Salber and Coutaz, 1993). It allows a prototype interface to be subjected to usability testing to ensure that the interface is understandable and appropriate for the task (Nielsen, 1994). By separating out the parts of a system that have not yet been developed and testing them separately in a WOZ type scenario, it is possible to define more precisely how these missing pieces should be developed.

The participants in a WOZ study do not know that a human is involved in running the system. Therefore, they are instructed to interact with the system as if it were fully operational. In this manner, it is possible to test how they would interact with a system without having to develop the system beforehand. This technique is quite useful in reducing the development time for interaction technologies and providing insight into how the system should be designed to maximize the interaction for the given task.

Dahlback et al. (1993) give a good overview of WOZ studies and how to design a quality study. They recommend the use of a cover story for a WOZ study to provide the participant with an objective to complete, so their focus will not be on the system itself but on the completion of the assigned task. This design encourages the user to interact with the system in a normal manner.

The participants must also be able to interact freely. Therefore, participants should not be told how to complete a task specifically as that influences their behaviour and restricts the interactions they may exercise. By using a convincing cover story, the users will be focused on the task at hand and not on how to use the system, thus eliciting more natural communication be-

haviour. However, the designers of a WOZ study should also be careful not to give their participants too much instruction, as doing so inhibits the natural communication techniques the users might use.

It is thus very important that during a WOZ study the participants think they are actually interacting with a working system (Dahlback et al., 1993). The human wizard cannot make simple errors, otherwise the participants will no longer believe they are working with an operational system. Therefore, care must be taken so the participants believe the actions and verbal responses are coming from the machine interface and not a person.

The behaviour of the wizard must also be consistent. For example, a given command from the user must always trigger the same behaviour from the wizard (Salber and Coutaz, 1993). The wizard must also react in an amount of time that is expected by the user. If the wizard is too slow to react, then the user may avoid using simulated functions believing they are not implemented or that the system is overloaded (Salber and Coutaz, 1993). The results from such a study aid in the development of the system by providing an indication of how participants would use the system in reality without having to implement it first, as well as how they feel about its implementation.

For example, Makela et al. (2001) found their WOZ study to be instrumental in the iterative development of their Doorman system. The Doorman is used to control the access of visitors and staff to their building and also to guide visitors upon entry into the building. Their WOZ study was designed such that the human wizard completed the speech recognition while the remainder of the system operated normally. In line with Dahlback et al. (1993), the interface for the wizard was kept simple to ensure a minimized response time and to reduce the possibility of errors. From this study, they found that they needed to shorten the utterances from the system to reduce communication time, provide the user with feedback to confirm that the system is operational and have better error handling.

To find out what kind of speech would be used with a robot in grasping tasks, Ralph and Moussa (2005) also conducted a WOZ study. In this study, users were asked to verbally instruct a robot to pick up five different small household items. The robot was fixed on a table and the users sat next to the robot while giving it instructions. Users were given a description of a primitive

command set and were allowed to modify these commands as the study progressed. The participants were asked to be as descriptive as possible in their commands. These commands were then translated into robot movement by the human wizard.

Results from the study showed that the participants felt natural language was an easy way to communicate with the robotic system. All participants were able to complete the pick and place tasks given to them. Participants tended to use short commands, in a mechanical manner, and mentioned that the interface would have been better had there been feedback from the robot once a command was given so the user would know that the command had been understood.

A WOZ experiment was employed by Carbini et al. (2006) for a collaborative multimodal story telling task. The objective of the study was to determine what speech and gestures would be used as two participants collaborated remotely with the system to create a story. In this study, the human wizard completed the commands of the user's speech and laser pointing gestures.

The instructions given to the participants were intentionally vague so as not to inhibit the actions of the users. This design is in line with Dahlback et al. (1993) who suggest using scenarios in WOZ studies and to not tell participants what to say or do explicitly as this inhibits what they would normally do in a given circumstance. Users in this study were found to complete a laser pointing gesture with a verbal command and they tended to point without stretching their arms.

Huettenrauch et al. (2006) conducted a WOZ study to determine spatial distance and orientation when a person is interacting with a mobile robot in a follow-me scenario. Two wizards controlled the mobile robot, one for robot navigation and on-board camera control and one wizard for spoken dialog. Although the intent of this study was not to interpret multimodal interaction, it did show that the use of the WOZ experimental design allowed the researchers to determine what is the best way for the robot to interact spatially with a human without having the complete system operational.

In their pilot WOZ study, Perzanowski et al. (2003) focused on verbal communication and gestural input through a touch screen to collaborate with a

remotely located mobile robot. They were specifically interested in finding out how people referred to objects when giving directions and trying to maneuver a mobile robot. Participants in this study were told they could talk to the robot as if it were human. In addition, they could point to objects and locations on a touch screen that included ego and exo-centric viewpoints. The participants were told to direct the robot to find an object. Little instruction was given to the participants, so as not to restrict their natural actions.

Two wizards interpreted the speech and touch gestures and drove the robot where they interpreted the user wanted the robot to go. They also spoke for the robotic system. Results from the study revealed that the users felt they had to continually guide the robot and therefore used a lot of short spoken commands. If the users had felt the robot was more autonomous they may have used more complex speech.

Overall, this brief review provides the following design principles for an effective WOZ study:

- The participants must be unaware that a wizard is involved.
- The behaviour of the wizard must be consistent and error free.
- A cover story should be used to focus the attention of the participants on completing a task and not on how they interact with the system.
- Participants should not be influenced on how they interact with the system.
- Participants must believe that they are interacting with a system that is fully functional.

3.4 Design Guidelines

Given the review of the general state of human-robot collaboration from the previous chapter, and the presentation and review of AR and its potential to enhance this type of collaboration, the design concepts behind the AR-HRC system can now be examined. Two important concepts must be kept in mind when designing an effective human-robot collaboration system. One,

the robotic system must be able to provide feedback as to its understanding of the situation and its actions (Scholtz, 2002). Two, an effective human-robot system must provide mechanisms to enable the human and the robotic system to communicate effectively (Fong et al., 2006).

The use of humour and emotion will enable the robotic agents to communicate in a more natural and effective manner, and therefore should be incorporated into the dialog management system. An example of the effectiveness of this type of communication can be seen in Rea, a computer generated human-like real estate agent (Cassell et al., 1999). Rea is capable of multimodal input and output using verbal and non-verbal communication cues to actively participate in a conversation.

The robot will need to understand the use of objects by its human counterpart, such as using an object to point or make a gesture. AR supports this type of interaction by enabling the human to point to a virtual object that both the robot and human refer to and use natural dialog such as "go to this point", thereby reaching common ground and maintaining situation awareness. In a similar manner, the robot would be able to express its intentions and beliefs by showing through the 3D overlays what its internal state, plans and understanding of the situation are. Thus, using the shared AR environment as an effective spatial communication tool.

Referencing a shared 3D environment will support the use of common and shared frames of references, thus affording the ability to effectively communicate in a truly spatial manner. As an example, if a robot did not fully understand a verbal command, it would be able to make use of the shared 3D environment to clearly portray to its collaborators what was not understood, what further information is needed, and what the autonomous agent believes could be the correct action to take. Or it could simply use that space to convey that it did not understand.

With the limited speech ability of robotic systems, visual cues will provide an important means of grounding communication. AR, with its ability to provide ego and exo-centric views and to seamlessly transition from reality to virtuality, can provide robotic systems with a robust manner in which to ground communication and allow human collaborative partners to understand the intention of the robotic system. AR can also transmit spatial awareness

though the ability to provide rich spatial cues, ego- and exo-centric points of view, and also by seamlessly transitioning from the real world to an immersive VR world. Therefore, the use of AR will enable the human to feel as if they are working side by side with a remotely located robot, providing a feeling of telepresence.

AR is an optimal method of displaying information. Billinghurst et al. (1998) showed through user tests that spatial displays in a wearable computing environment were more intuitive and resulted in significantly increased performance. Figure 3.15 shows spatial information displayed in a head stabilised and body stabilised fashion. Using AR to display information, such as robot state, progress and even intent, will enhance understanding, grounding, and thus collaboration.

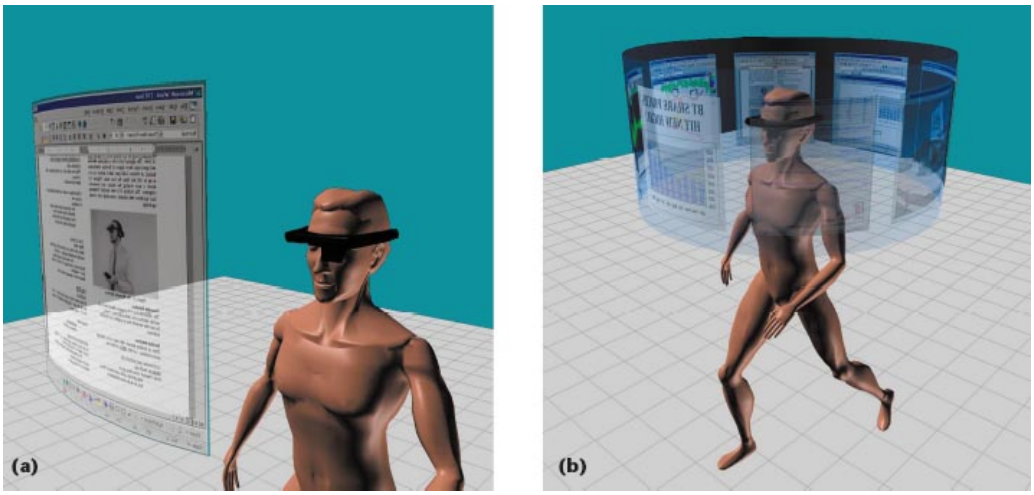


Figure 3.15 Head stabilised AR display (a) and body stabilised (b) (Billinghurst et al., 1998).

Humans and robots have different strengths. To create an effective human-robot collaborative team, the strengths of each member will need to be capitalized on. Humans are good at using vague spatial references. For example, most people would point out where an object is by using some sort of deictic reference, such as "it's over there". Unfortunately, robotic systems are not designed to understand these ambiguous references. Therefore, for a human-robot collaboration system to be natural to a human it will have to be able to understand vague spatial references. Similarly, for a collaboration system to be effective for robotic systems it will have to translate vague spatial references into exact spatial coordinates that a robotic system needs to operate.

Humans are good at dealing with unexpected and changing situations. Robotic systems, for the most part, are not. Robots are good at physical repetitive tasks that can tire human team members. Robots can also be sent into dangerous environments in which humans cannot work. Therefore, for a human-robot team to collaborate at the most effective level the system should allow for varying levels of autonomy enabling robots to do what they do best and humans to do what they do best.

By varying the level of autonomy the system would enable the strengths of both the robot and the human to be maximized. Varying levels of autonomy would allow the system to optimize the problem solving skills of a human and effectively balance that with the speed and physical dexterity of a robotic system. Adjustable autonomy enables the robotic system to better cope with unexpected events, being able to ask its human team member for help when necessary.

For robots to be effective partners, they will need to interact meaningfully through mutual understanding. A robotic system will be better understood and accepted if its communication behaviour emulates that of humans. Communication cues should be used to help identify the focus of attention, greatly improving performance in collaborative work.

Grounding is an essential component of communication and is reached through meaningful interaction and the exchange of dialog. The use of humour and emotion can increase the effectiveness of a robot to communicate, just as in humans. Communication cues, such as the use of humour, emotion, and non-verbal cues, are essential to communication and thus, effective collaboration. Therefore, it is evident that for a human-robot team to communicate effectively, all participants will have to feel confident that common ground is easily reached.

The ability to maintain situation awareness is another important aspect of effective collaboration. Studies have shown that the lack of situation awareness has detrimental effects on human-robot team performance. Therefore, a human-robot collaboration system should provide the means for both human and robotic team members to maintain situation awareness.

Reference frames are fundamental if a human team member is going to

use spatial dialog when communicating with robotic systems. Consequently, a robust collaborative system should effectively allow for human team members to use reference frames at will and translate this information into a usable format for the robotic system. The collaborative system should enable a robot to be aware of its surroundings and the interaction of collaborative partners within those surroundings. If a human-collaboration system entails these design parameters, the result should be an effective natural collaboration between humans and robotic systems.

Finally, research has shown that a multimodal interface leads to a more natural experience for the user. A speech and gesture interface is more powerful than either modality alone. Multimodal interfaces are more intuitive because they combine the strengths of the different modalities. For these reasons, a multimodal approach has been taken in the design and implementation of the AR-HRC system.

3.5 Summary

The research in this thesis attempts to fill a gap in current HRI research by providing a platform that incorporates the following:

- Natural human-robot communication based on concept of grounding
- Use of real world objects and gesture interaction
- Referencing of objects and points in 3D space through common frame of reference provided by AR environment
- Increased situation awareness through a shared work space
- Effective transmission of robot internal state and intention
- Use of adjustable autonomy to capitalize on strengths of all team members
- Ability for robot to interact meaningfully with human
- Collaborative environment for HRI

- Projection of the human into the work space of a remotely located robot, telepresence
- Formal user study of AR interface for HRI

Such a system offers the potential to provide the necessary tools to optimize human-robot collaboration.

Chapter 4

Multimodal AR Interaction

Chapter 2 reviewed the state of the art in HRI while Chapter 3 introduced AR and outlined its benefits. Chapter 3 also provided a discussion on how AR can provide the required environment for an effective human-robot collaboration system. Multimodal interfaces were investigated and found to be a powerful mechanism for human-machine interaction. Therefore, a multimodal interface is an optimal design choice for the AR-HRC system.

This chapter discusses the development of a multimodal AR application. Lessons learned from this development were incorporated into the design of an AR application for human-robot collaboration, which will be covered in detail in Chapter 5.

4.1 Architecture

To investigate multimodal interaction in an AR environment, a system was developed that used speech and paddle based gestures as inputs to an AR system. The system developed is a modified version of the VOMAR application for tangible manipulation of virtual furniture in an AR setting (Kato et al., 2000). The VOMAR application used a single input device, a handheld paddle, to manipulate virtual objects. The paddle can be seen in Figure 4.1.

The user was able to perform various tasks by using real world paddle movements that were mapped to virtual object motion in the AR setting. The application domain for this work was interior design. Therefore, the VOMAR application allowed the user to arrange virtual furniture in a virtual room.

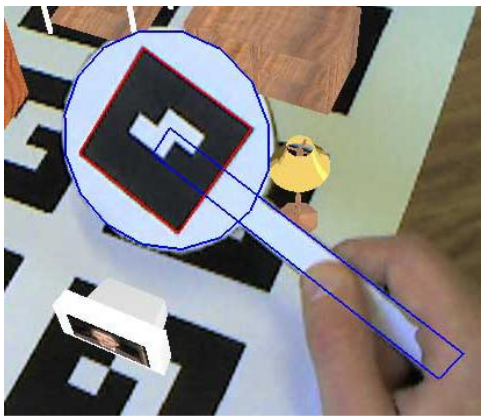


Figure 4.1 The single input modality handheld paddle for the VOMAR application. The real world paddle is held by the user. The paddle is outlined in blue in AR.

The original VOMAR application used only a single modality; paddle input. In the work described in this chapter, speech input was incorporated as well to create a multimodal system that understood combined speech and paddle-based gesture input. The result was the Multimodal Augmented Reality System (MARS).

The MARS architecture is shown in Figure 4.2. As can be seen, it is made up of several modules. The speech processing module is responsible for recognizing the spoken dialog of the user. It is also responsible for the text to speech (TTS) output of the application, which is the verbal response to the user. The Dialog Management System (DMS) module compares the spoken dialog that the speech processor recognized with predefined goals for the system. The speech processor and DMS make use of the Ariadne spoken dialog system (Denecke, 2002).

When a goal has been reached through spoken dialog, the DMS sends the appropriate command to the AR module via the Multimodal Communication Processor (MCP). The MCP is built upon the Internet Communications Engine (ICE) (ZeroC, 2008). If the command sent is to be combined with a gesture, the Augmented Reality (AR) module checks if the paddle is in the users view and then calculates its position.

The AR module also simultaneously calculates the location of all virtual objects, whether they are on the menu pages or in the virtual room, and compares these locations to that of the paddle. When the paddle is within

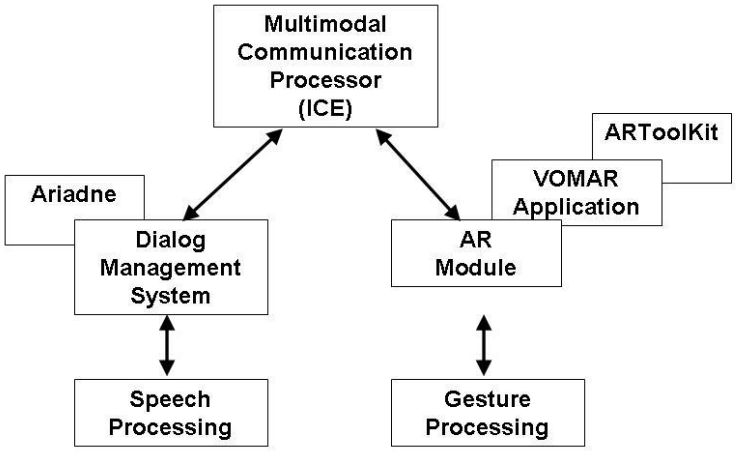


Figure 4.2 System architecture for the MARS application.

a predefined distance of a virtual piece of furniture, this piece of furniture becomes active and the verbal command sent in from the DMS is applied to the selected piece of furniture. The AR module is written in C++ and uses the VOMAR (Kato et al., 2000) and ARToolKit (ARToolKit, 2008) libraries.

A user may combine speech commands and paddle gestures to interact with the system. To understand the combined speech and gesture interaction, the system must fuse input from both of these streams into a single discernible command. This fusion is achieved by recording the time each input occurs through the use of an event time stamp. Therefore, the paddle and speech input can be considered for fusion only if the input time stamps from both input streams are within a certain time frame of each other. This time window was set to five seconds as this amount of time was deemed sufficient to allow the user to select an object and not allow a command to remain active for too long.

4.2 MARS Application

The objective of the MARS application was to allow people to easily and effectively arrange AR content using a natural mixture of speech and gesture input. A single fiducial marker located on the end of the paddle is used for gesture input. This fiducial marker allows the paddle's position and orientation

in the AR environment to be determined by using the ARToolKit (ARToolKit, 2008) library. This approach ensures a measure of computational and user simplicity.

A4 sized pages containing an array of fiducial markers serve as menu pages holding virtual furniture models, as shown in Figure 4.3. As the user looks at each of the A4 container pages through a Head Mounted Display (HMD), they see different sets of virtual furniture. The 3D virtual models appear superimposed over the real pages aligned with the fiducial markers. The square fiducial markers printed on these pages are used by the ARToolKit library to locate and place the virtual content.

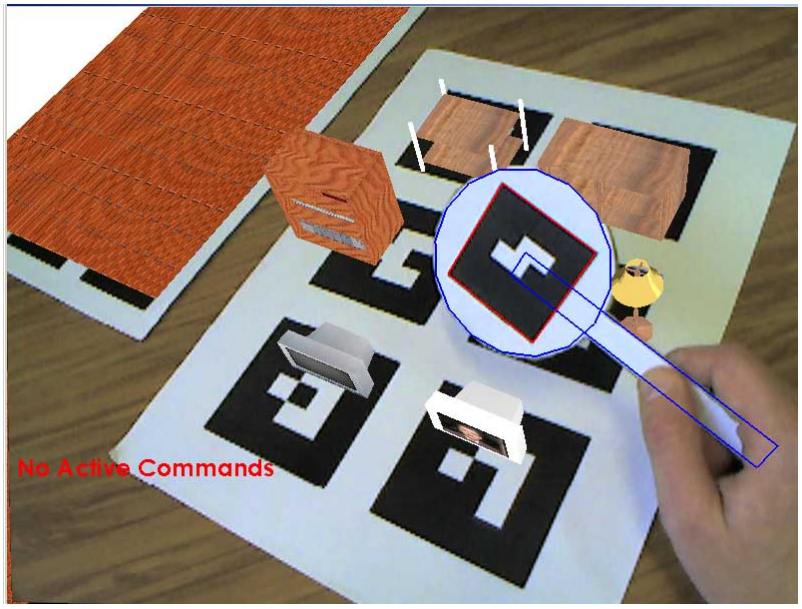


Figure 4.3 An A4 sized sheet with an array of square fiducial markers is used as a menu page containing various pieces of furniture. The user is shown holding the real world paddle that is used to interact with the virtual content.

A separate larger sheet also containing AR fiducial markers serves as the workspace. This page displays a 3D graphic of an empty room where the virtual models of furniture are to be placed. Furniture can be arranged in the room by selecting various pieces of virtual furniture and placing them in the virtual room, as shown in Figure 4.4.

Looking at the workspace page for the first time the user sees an empty virtual room. The user is then able to transfer objects from the menu pages to the virtual room using paddle and speech commands. Figure 4.5 shows the

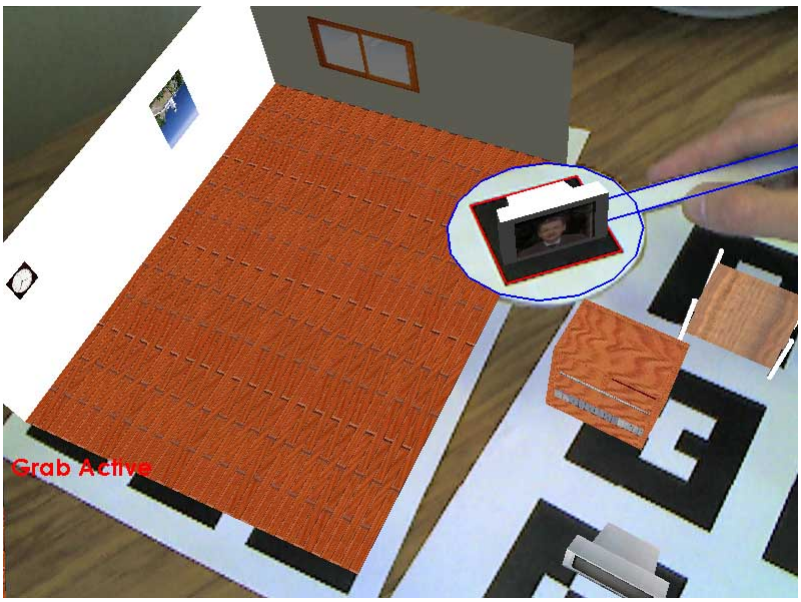


Figure 4.4 A separate sheet containing virtual markers locates the virtual room. Shown here is the view the user sees with the room initially empty and a virtual object on the real world paddle ready to be placed in the room.

result of the multimodal interaction for virtual object manipulation. The user can also modify the position and orientation of furniture already in the virtual room through the use of the real world paddle and speech input.

The system provides visual and audio feedback to the user. The speech interpretation result is shown as a text overlay on the screen and audio feedback is provided after the speech and paddle gesture commands are recognized. Therefore, the user is immediately notified when the system recognizes speech or gesture input.

4.3 The Modalities

The MARS application is capable of working in three different modes, a gesture only mode, speech input with static paddle placement, and full paddle gesture combined with speech. The gesture only mode works essentially the same as the initial VOMAR application. In this mode, the user interacts with the system through paddle commands only and, as a result no explicit fusion strategy is needed. In this section each of the modalities is described in more detail.

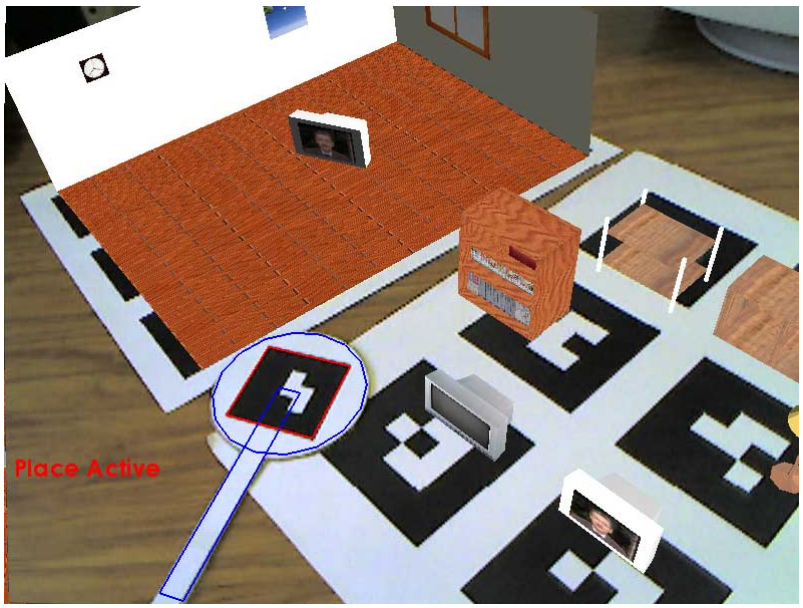


Figure 4.5 Virtual object placed in room where paddle was located when verbal placement command was issued.

4.3.1 Gesture Only

The gesture only mode consists of a variety of interaction techniques for object selection, manipulation and deletion. For example, for manipulation, the user is able to select a piece of virtual furniture from the menu page with an empty paddle by holding the paddle over the object. If the paddle is empty and placed on a virtual object, the object is copied onto the paddle after it has remained in this position for a short period of time

Other commands are also supported. If there is an object attached to the paddle, when the paddle is tilted the object will slide off the paddle and into the empty virtual room. If there is an object attached to the paddle, the object will be deleted from the paddle when the paddle is shaken from left to right.

An object already placed in the room can be moved around by "pushing" it with the paddle. An item of furniture that is placed in the room can be deleted by hitting it with the paddle. The paddle gestures are recognized by a gesture input library as part of the original VOMAR application. Object manipulation through the use of the handheld paddle is provide in Table 4.1.

Once an object is on the paddle, it can be picked up and viewed from any viewpoint. These interactions are very natural to perform with the real paddle,

Command	Paddle Gesture
Pick	Hold paddle next to object
Place	Tilt paddle to slide virtual object off
Move	Push object with paddle
Delete (object on paddle)	Shake paddle left to right
Delete (object in room)	Tap object with paddle

Table 4.1 Handheld paddle commands for MARS application.

so in a short period of time a user can assemble a fairly complex arrangement of virtual furniture. However, placement of the virtual furniture is this manner is not very precise due to errors in the ARToolKit tracking and other factors.

4.3.2 Speech with Static Paddle Gestures

A second mode of interaction is to use speech combined with static paddle placement. In this mode, the user interacts with the virtual content using speech and paddle placement. However, the system only considers the static paddle pose at a particular time and fuses this information with the speech recognition result to interpret the combined speech and gesture commands. Therefore, the paddle was only used for object selection.

This mode works in the following manner. When a speech command is recognized, it is checked against a set of goals. If a match is found, the appropriate command id number is sent to the AR application, where it is acted upon.

For example, consider the speech input "grab this" while the user has placed the paddle on a virtual object on one of the menu pages. The system will check the position of the paddle and compare it to the other virtual objects in the scene. If the paddle is within a predefined threshold of an object, then that object will be selected.

In this example, the object is grabbed, or selected from the contents page and placed on the paddle for further action. If the paddle position is not within the predefined threshold, then no object will be acted upon. Figure 4.6 shows the process flow to recognize and act upon a speech command.

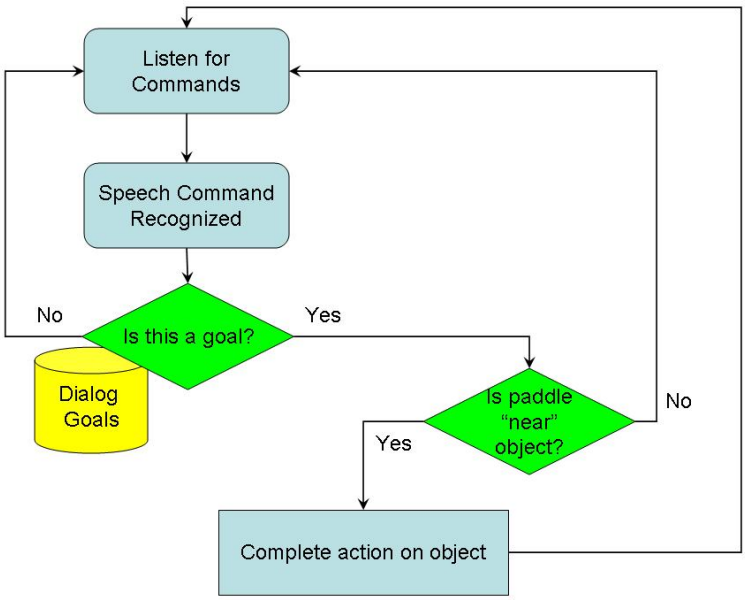


Figure 4.6 Process flow to recognize and act on speech command.

4.3.3 Speech and Gesture

The final mode of interaction is full paddle gesture combined with speech. The user is able to interact with the system using both speech and continuous paddle gestures. This mode is a combination of the two modes explained previously, but this time continuous paddle input is used.

For example, the user can give a speech command "grab this" to select an object. The object can then be placed in the virtual room using the paddle tilting gesture, which was not available in the static paddle mode of the previous section. In this manner, the user can easily combine speech and paddle gesture input and choose which interaction technique is more appropriate for the given task.

When the system recognizes a speech command and matches it to a dialog goal, a time stamp is recorded. The system then determines if the paddle is within a defined threshold of any of the actionable objects. If within five seconds the paddle is found to be close enough to an object, then that object is acted upon. However, if no item is found within five seconds, then the command expires and the user must issue any further desired command.

4.4 Speech Commands

A list of speech commands the system can process are given below:

- **Delete Command:** This command will delete an object from the paddle or from the workspace area. If there is an object on the paddle, it will be deleted. If there is no object on the paddle and the workspace is in view, the object the paddle is touching will be deleted. An example of such a command is *"delete this"*.
- **Translate Command:** If the workspace is in view, this command attaches a virtual object in the workspace to the paddle so that it follows the paddle translation. The object will be released from the paddle after the user gives the *Stop* or *Place* command. The translate command would be *"translate this"*.
- **Rotate Command:** This command has a similar function as the Translate command. It attaches a virtual object in the workspace to the paddle so that it can follow the paddle rotation. The object will be released from the paddle after the user gives the *Stop* or *Place* command. An example of this command is *"rotate that"*.
- **Move Command:** This command combines the Translate and Rotate commands. It attaches a virtual object from the workspace to the paddle so that it follows both paddle translation and rotation. The object will be released from the paddle after the user gives the *Stop* or *Place* command. The user would issue the following command *"move this"*.
- **Place Command:** If there is an object attached to the paddle, this command places the attached object at the paddle location in the workspace. An example of this command is *"place here"*.
- **Stop Command:** This resets a *Delete, Translate, Rotate* or *Move* command. This command is the single word *"stop"*.

4.5 Evaluation

The MARS application was demonstrated at the International Conference on Artificial Reality and Telexistence (ICAT2005). The system was set up using

a standard desktop PC. Users wore a head mounted display (HMD) with a web cam aligned with the user's line of sight. Marker grids were laid out in front of the user so that it was easy for the users to select objects from the menu pages and place them in the virtual room. A list of verbal commands was placed in front of the user for reference.

Users were quickly able to manipulate the virtual content in an effective manner and commented that the system was easy to use. The system was able to understand native and non-native English speakers from various countries. A user trying out the system during this demonstration can be seen in Figure 4.7.

Figure 4.7 Demonstration of MARS application at ICAT2005 conference.

In addition, a formal user evaluation study was conducted to determine if the multimodal interface actually improved the efficiency of user interaction in the AR environment (Irawati et al., 2006). The set up for the user study was similar to that for the ICAT2005 demo described previously. In the study, participants were to build a predefined arrangement of furniture using the three modalities provided by the system:

- Paddle Gestures Only
- Speech with Static Paddle Gestures
- Speech with Paddle Gestures

To minimize order effects, the presentation sequences of the three interface conditions and three furniture configurations were systematically varied between users. Before each trial, a brief introduction and demonstration was given so that the participants could become familiar with the system. Participants were also allowed to practice until they were proficient enough in the given condition to assemble a sample scene in less than five minutes. A list of speech commands was provided for reference during the experiment.

Participants completed the task significantly faster using the speech and static paddle condition. This result shows that the use of multiple input channels leads to an improvement in task completion time. Subjective questioning showed that users felt they completed the tasks more efficiently when using the multimodal interface. A complete discussion of this user study can be found in (Irawati et al., 2006).

4.6 Summary

In this chapter, the development of a multimodal AR (MARS) application was described. The MARS application is utilized in this research as a test bed to determine the effectiveness of multimodal interaction in an AR environment. A formal evaluation of the interface showed that combining speech and paddle gestures improved performance over the use of a single modality alone. These positive study outcomes provided the impetus to incorporate this same type of interaction into the AR-HRC system.

Chapter 5

Architectural Design

Chapter 4 discussed the development of a multimodal AR system and how the performance of that system was improved through the use of a multimodal interface. As a result of that work, a multimodal approach has been taken in the design and development of the AR-HRC system in this thesis.

This chapter first provides an overview of the design approach taken for the AR-HRC system and then describes each module in detail. Two examples are provided to help define how the different modules interact and how the goals of the AR-HRC system are achieved. A particular focus is placed on highlighting how each module contributes to the overall multimodal nature of the system.

5.1 Design Approach

A number of the capabilities required for a robust human-robot collaboration system were identified in Chapters 2 and 3 through a review of research in the areas of communication, HRI and AR. This section lists these capabilities individually and details how they are incorporated in the design of the AR-HRC system presented.

Communication is a fundamental part of collaboration. Therefore, the communication link between the robot and human must be as robust as possible. However, communication is more than just verbal exchanges. Therefore, a multimodal approach has been taken in the design of the AR-HRC system which ensures that the communication is as robust as possible.

This design enables the human to communicate in a multimodal fashion through the use of speech and paddle driven gestures. The robotic system is also able to communicate in a multimodal fashion through the use of synthesized speech and virtual graphic overlays. Integration of AR technology into the system design is a key factor that enables both the human and robot to communicate effectively. More specifically, AR enables both the human and robotic system to communicate in a multimodal fashion.

Natural communication for the robot has been integrated into the system through the use of random verbal responses and humour. To avoid the robot uttering the same phrase for a given context, the system randomly selects one of a number of responses that is appropriate. Therefore, the robotic system is able to communicate in what is perceived by the human as a natural manner, since the robot is not repeating itself verbatim, but is using a variety of phrases for a given situation. Hence, it should elicit natural communication behaviour from the human in return.

In any truly collaborative effort, the participants must be able to easily reach common ground. Without being able to reach common ground, conversation partners will not be able to communicate effectively. Effective communication is even more important in human-robot communication since it is more difficult to repair dialog when a misunderstanding occurs. If the human is unable to easily reach common ground with the robotic system, then collaboration is hindered and the human will lose confidence in the system.

The design of the AR-HRC system enables grounding for robotic communication. This grounding takes place through a combination of verbal and visual feedback to the user. In essence, this approach is similar to a human using a verbal reference and clarifying that reference with a gesture.

The AR-HRC system allows the robot to respond with a verbal statement and follow that statement with a visual overlay in the AR environment. This method enables the human to easily reach common ground with what the robot has "said" and closely follows the way humans communicate, thus providing the human with a familiar form of communication.

For a human-robot collaborative system to be truly effective, it should also incorporate adjustable autonomy, which is the ability for the robotic system to

vary the level of human input required during execution. Fundamentally, this means that the robot can operate at the level of autonomy that is appropriate for the given situation. The human is able to monitor the robotic system and get involved if warranted. On the robot side, the robot works autonomously once a plan has been interactively created and reviewed. However, during execution, the robot can ask the human for help if a situation arises where the robot cannot find an optimal solution itself.

The effect of this system design is that the robot is able to do what it is best at, repetitive physical tasks in hazardous environments. A human also offers the collaborative team their capabilities of dealing with unexpected situations. Therefore, the system is designed to capitalize on the strengths of all members of the collaborative team effort.

An effective collaborative team needs to maintain awareness of what the various members of the team are doing and how actions of any member will affect the team as a whole. This process is called maintaining situation awareness. Maintaining situation awareness is very important in a human-robot collaborative effort as the robot may not be able to indicate to the human partner every detail about its surroundings. In the AR-HRC system design, the human is able to maintain situation awareness of the robot by having the 3D visual representation of the robot in its environment overlaid onto the real world view of the human. This overlaid view is accomplished through the implementation of the AR module.

The AR-HRC system architectural design is shown in Figure 5.1. The following sections describe each module of the design in detail. A particular focus is also paid to how each module specifically contributes to the design goals outlined.

5.1.1 Speech Processing

Speech processing consists of recognizing and parsing human speech. Input for speech processing is through a noise canceling microphone worn by the user. The user's spoken input is converted into a text string using the Microsoft Speech Engine (MicrosoftSpeech, 2007). The parsed speech is then compared to a set of defined dialog goals and used to initiate system commands. If a goal

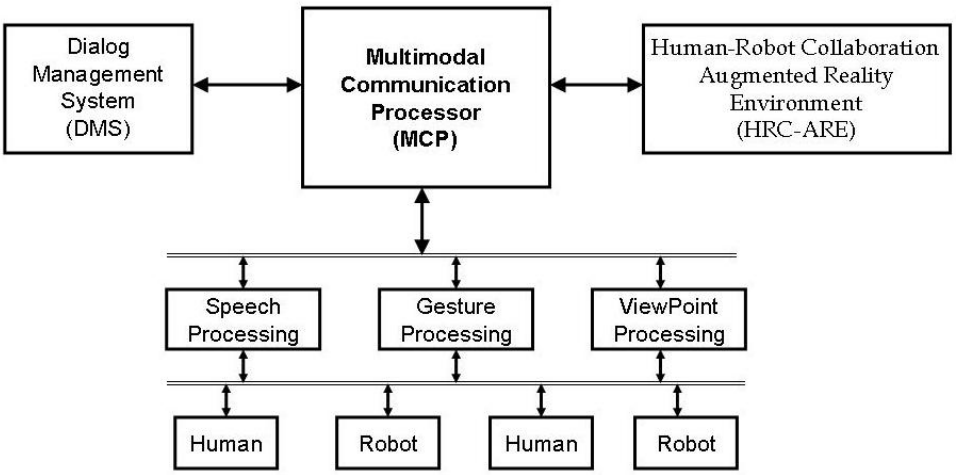


Figure 5.1 The AR-HRC System Architecture.

is matched then this information is sent to the Multimodal Communication Processor (MCP) module, which is discussed in Section 5.1.6.

Dialog goals are specified using an xml based rule description. An example of the defined dialog goal for the "go here" command is shown in Figure 5.2. The "RULE_ID" indicates that the "go here" command is a move command. The first item "please" is optional, by using the <O> tags, thus the system will recognize the "go here" command whether or not the human precedes this command with "please". The human can then use either "move" or "go" to initiate the command. The "RULEREF MoveDeictic" item shows that to complete the dialog the human can use either "here" or "there". So valid commands include "please move there", "move here", "please go here", "go there" etc.

In this manner, the dialog understood by the system is flexible and adaptable to different users. The same command information is sent to the MCP regardless of which variation of the dialog goal the user decides to use. The example provided in Figure 5.2 can easily be modified to expand the dialog that the AR-HRC system understands, making it compatible with a large number of users.

Another input for speech processing comes from the MCP. This input is the spoken dialog of the robotic system. Using the text-to-speech capabilities

```
# Top level rule for "go here" verbal command
<RULE ID="VID_Move" TOPLEVEL="ACTIVE">
<O>Please</O>
        <P>
            <L>
                <P>move</P>
                <P>go</P>
            </L>
        </P>
        <P>
      <L>
        <RULEREF REFID="VID_MoveDeictic" />
      </L>
    </P>
</RULE>

# Spatial references
<RULE ID="VID_MoveDeictic" >
   <L PROPID="VID_MoveDeictic">
      <P VAL="VID_MoveDeicticLocale">here</P>
      <P VAL="VID_MoveDeicticLocale">there</P>
   </L>
</RULE>
```

Figure 5.2 Speech: Processing of "go here" command.

of Microsoft Speech (MicrosoftSpeech, 2007), this information is presented to the user in the form of verbal output through a set of speakers connected to the system. The dialog spoken by the system is stored as a list of strings. When a situation occurs where the system needs to speak, it selects from a variety of strings that are defined for the given situation. Therefore, the system communicates in a more natural manner by using a variety of phrases for each situation.

5.1.2 Dialog Management System

The Dialog Management System (DMS) takes input from the MCP. This input is the defined goal reached through speech processing. The DMS matches this defined goal to an action for the system to take. This output is then sent to the MCP for processing. To continue with the "go here" example of Section 5.1.1, Figure 5.3 shows the matching of the defined speech goal to an action for the system to take.

The "if (SUCCEEDED)" command checks to see if a dialog goal, from Section 5.1.1, has been reached. If a dialog goal has been reached, then this

```
if (SUCCEEDED(pPhrase->GetPhrase(&pElements)))
{
  switch(pElements->Rule.ulId)
  {
    case VID_Move:
      switch(pElements->pProperties->vValue.ullVal)
      {
          case VID_MoveDeicticLocale:
             cout << "Move here/there" << endl;
             commandToSend = "hrcAre go here";
             break;

          // .. other cases
      }

    // ... other cases
  }
}
```

Figure 5.3 DMS: Processing of "go here" command.

goal is identified through the switch statement that follows. In this case, a move command has been identified. A second switch statement completes the command by identifying where the move command is directed. Once the completion of the goal has been defined, a command is sent to the appropriate module to take action. In this example, the command is sent to the gesture processing module, discussed in Section 5.1.3, to define the point referred to in the deictic reference "here" or "there" and complete the multimodal command.

The MARS application described in Chapter 4 made use of the open source Ariadne spoken dialog system (Denecke, 2002). However, for the AR-HRC system the DMS and speech processing modules were developed and integrated for the specific needs of the system. These modules are built on the Microsoft Speech SAPI 5.1 (MicrosoftSpeech, 2007) software libraries.

5.1.3 Gesture Processing

The human is able to gesture to the robotic system through the use of a paddle. This paddle has a fiducial marker on it that allows the ARToolKit (ARToolKit, 2008) to track its position. When a dialog goal is reached that requires a gesture to complete the command, the MCP sends a request that the position and orientation of the paddle be recorded and used as the 3D point in space to complete the verbal command.

The coordinates of the point are translated into the coordinate frame of the robot. Therefore, the user is able to use a generic spatial reference, such as "go here", and gesture into the 3D virtual world of the robot with the paddle selecting a point in 3D space. This generic communication mechanism is then transferred into precise coordinate information that the robot needs to execute the maneuver. Figure 5.4 shows how gesture processing takes place for the "go here" example.

Once the DMS module of Section 5.1.2 has identified a command that requires completion through gesture interaction, a command is sent to the gesture processing module to find the paddle position and the gesture parameters. The process begins by first making checks to see if the paddle is in view, then to make sure the paddle is in pointer mode and finally, if the tracking grid is visible that defines the virtual work space of the remote robotic system. The system then calculates the coordinates of the paddle in relation to the remote world environment.

These coordinates are then transferred into the reference frame of the robot. A check is made to ensure that the coordinates are within the working boundaries of the robot. If this check passes, then the type of command is defined through the following "if" statement where the variable "action" contains the command type. This variable is set by the input from the DMS module.

At this point, a command has been completely defined. A verbal response is sent to the user via the "sayThisRandom" method and then the coordinates selected are recorded as the position to proceed to. The "drawTrajectory" variable is set to true, telling the system to draw this new trajectory as an overlay in the AR environment. Therefore, the user is able to immediately verify that the robot has understood the 3D point selected and can see how the robot plans to proceed to this point. The robot does not begin to move until the user tells the robot to "execute the plan".

5.1.4 Viewpoint Processing

The viewpoint-processing module interprets the user's viewpoint through the use of a camera attached to a head mounted display (HMD) worn by the user.

```
  //  Check if paddle is visible
  if (seePaddle)
  {
    //  Make sure paddle mode is pointing
    if (paddleAsPointer)
    {
      // Make sure can see robot's world
      if (baseMarker->isValid())
      {
        // Find position of paddle in world coordinates
        osg::Matrixd* paddleNodeMatrix = getWorldCoords(paddleGroup);
        osg::Vec3f paddleWorldCoords = paddleNodeMatrix->getTrans();

        //  Find coordinates of robot wold
        osg::Matrixd baseMatrixFindCoords = baseMarker->getTransform();
        baseMatrixFindCoords.invert(baseMatrixFindCoords);
        osg::Vec3f baseMatrixWorldCoords = baseMatrixFindCoords.getTrans();

        *paddleNodeMatrix = (*paddleNodeMatrix) * (baseMatrixFindCoords);

        //  Find paddle position in reference frame of robot's world
        osg::Vec3f paddleFromBase = Utils::distancePaddleFromBase
        (baseMarker->getTransform(), paddleMarker->getTransform());

        //  Make sure selected point is valid
        bool selectedPointOK = Planner::checkCoordsValid(paddleNodeMatrix->getTrans(),
                                                         baseScale);

        if (selectedPointOK)
        {
          osg::Vec3f paddlePos = paddleNodeMatrix->getTrans();

          //  Check if have command and that it is "go here"
          if (haveCommand && action == HERE)
          {
            //  Respond to user with random verbal response.
            sayThisRandom(goHereResponse);

            //  Assign position of paddle as the go to position
            goToPosition(paddlePos);

            //  Tells system to draw trajectory with point selected in AR environment
            drawPlanTrajectory = true;
          }

          //  Other deictic command such as FRONT, BEHIND, LEFT and RIGHT

        }
      }
    }
  }
```

Figure 5.4 Gesture: Processing of "go here" command.

Computer vision algorithms in ARToolKit are used to calculate the position of the camera. Since this camera is aligned with the direction the user is looking in, the user's viewpoint direction into the 3D world is therefore acquired. The viewpoint direction of the user is then used to define the location the user is referring to when using a phrase, such as "go to the left of this", while selecting an object in the AR world with the paddle. As example of this type of processing is provided in Section 5.2.

5.1.5 HRC-ARE

The Human-Robot Collaboration Augmented Reality Environment (HRC-ARE) provides the user with a 3D virtual representation of the robot and its work environment. The HRC-ARE is built upon the osgART libraries (Looser et al., 2006). The osgART libraries use an Open Scene Graph (OpenSceneGraph, 2008) wrapper for the ARToolKit (ARToolKit, 2008) and were selected for use in the AR-HRC for their high level rapid prototyping approach to creating virtual content for AR environments.

5.1.6 Multimodal Communication Processor

The MCP is responsible for receiving information from the other modules and sending information to the appropriate modules. Thus, the MCP is responsible for combining multimodal input, registering this input into something the system can understand, and then sending the required information to other system modules for action. The result of this system design is that a human is able to use natural speech and gestures to collaborate with robotic systems.

5.2 Deeper Spatial Dialog

The previous sections described the components that make up the architectural design of the AR-HRC system. Using an example of the "go here" command the interaction between the modules was presented. This section attempts to go into more detail of how the gesture, speech and viewpoint information is handled. An example with deeper spatial dialog is used, namely that of the

human using the spatial reference “behind this”. This example will also be in the midst of collaboratively creating a plan for the robot. In particular, a via point will be added to a path whose creation is currently in progress. A via point is an intermediate point added to the path to ensure the robot avoids collisions and follows a desired trajectory.

The speech processor listens for a verbal command from the user. While listening the speech processor compares the verbal input to defined dialog goals. In this example, the user’s verbal command is “place via point behind this”. This defined dialog goal is shown in Figure 5.5 and shows how the user can select from various verbal commands to achieve the same dialog goal. The dialog goal can also be completed such as “place via point here”. Also shown is the defined dialog goals for other spatial references such as “left of”, “right of” and “in front of”. A description of how this code is interpreted was provided in Section 5.1.1.

Now that a dialog goal has been defined, the appropriate message is sent to the MCP. The MCP then sends this information to the DMS which matches the defined dialog goal to an action that needs to be taken by the system. Figure 5.6 shows how the dialog goal is matched to an action for the system. For an explanation of how the code works, please refer to Section 5.1.2.

The defined dialog has now been matched with an action. The HRC-ARE is now responsible for defining the location of the spatial reference. First, the HRC-ARE seeks to define what object the user is selecting. This task is done by comparing the location of the handheld paddle with the various objects in the robot’s environment. Figure 5.7 shows how an object is selected through paddle manipulation.

Checks are first made similar to those defined in Section 5.1.3 and shown in Figure 5.4. In the case of the spatial reference “behind this”, the system must determine what “this” refers to. The system calculates the position of each object in the work space and then compares these positions to that of the paddle. If the paddle is found to be within a proximity threshold of one of the objects, then this object is selected.

The system must now define what “behind” means in reference to the object selected with the paddle. To determine this reference, the system first

```
<RULE ID="VID_ViaPoint" TOPLEVEL="ACTIVE">
  <O>please</O>
    <P>
      <L>
        <P>place via point</P>
        <P>place middle point</P>
        <P>place way point</P>
      </L>
    </P>
    <P>
      <L>
        <RULEREF REFID="VID_MoveDeictic" />
        <RULEREF REFID="VID_MoveDirection" />
      </L>
    </P>
</RULE>

<RULE ID="VID_MoveDirection" >
  <L PROPID="VID_MoveDirection">
    <P VAL="VID_MoveDeicticLeftOf">to the left of this</P>
    <P VAL="VID_MoveDeicticLeftOf">to the left of that</P>
    <P VAL="VID_MoveDeicticRightOf">to the right of this</P>
    <P VAL="VID_MoveDeicticRightOf">to the right of that</P>
    <P VAL="VID_MoveDeicticFrontOf">in front of this</P>
    <P VAL="VID_MoveDeicticFrontOf">in front of that</P>
    <P VAL="VID_MoveDeicticBehind">behind this</P>
    <P VAL="VID_MoveDeicticBehind">behind that</P>
  </L>
</RULE>

<RULE ID="VID_MoveDeictic" >
  <L PROPID="VID_MoveDeictic">
    <P VAL="VID_MoveDeicticLocale">here</P>
    <P VAL="VID_MoveDeicticLocale">there</P>
  </L>
</RULE>
```

Figure 5.5 Speech: Processing of "place via point behind this" command.

```
if (SUCCEEDED(pPhrase->GetPhrase(&pElements)))
{
  switch(pElements->Rule.ulId)
  {
    case VID_ViaPoint:
      switch(pElements->pProperties->vValue.ulVal)
      {
        case VID_MoveDeicticBehind:
          cout << "Via point behind this/that" << endl;
          commandToSend = "hrcAre viapoint behind that";
          break;

          // .. other cases
      }

    // ... other cases
  }
}
```

Figure 5.6 DMS: Processing of "place via point behind this" command.

```
  //  Find object paddle comes near
  std::vector<osg::Node*>::iterator iter;
  bool foundOne = false;
  for (iter = trackableNodeList.begin(); iter != trackableNodeList.end(); iter++)
  {
    //  Find position information of individual item in trackableNodeList
    osg::Node* thisNode = *iter;
    osg::MatrixTransform* thisMatrixTransform = dynamic_cast<osg::MatrixTransform*>(*iter);

    //  Calculate world coordinates
    osg::Matrixd* nodeWorldCoords = getWorldCoords(thisNode);
    osg::Vec3f nodeWCoords = nodeWorldCoords->getTrans();

    //  Get world coordinates of paddle
    osg::Vec3f paddleNodeMatrixCoords = paddleNodeMatrix->getTrans();

    //  Calculate distance of paddle from base
    osg::Vec3f nodeCoordsBase = Utils::distancePaddleFromBase(baseMarker->getTransform(),
                                                              *nodeWorldCoords);

    //  Calculate the distance from paddle to individual trackable item
    float distancePaddleToCurrNode = (paddleNodeMatrixCoords - nodeWCoords).length();

    //  If paddle is close to item, select this item
    if (distancePaddleToCurrNode < PROXIMITYTHRESHOLD)
    {
      foundOne = true;
      findNodeVisitor findCubeColorSwitch("cubeColorSwitch");
      thisNode->accept(findCubeColorSwitch);

      //  Turn selected object green for visual user feedback
      osg::Switch* colorSwitch = dynamic_cast<osg::Switch*> (findCubeColorSwitch.getFirst());
      colorSwitch->setSingleChildOn(1);
    }
  }
```

Figure 5.7 Gesture: Processing of “place via point behind this” command.

calculates the user's viewpoint in relation to the object. The viewpoint is then used to calculate a point offset from the object selected in the direction defined by the spatial predicate, "behind" in this example. Figure 5.8 shows an example of how the spatial predicate is defined.

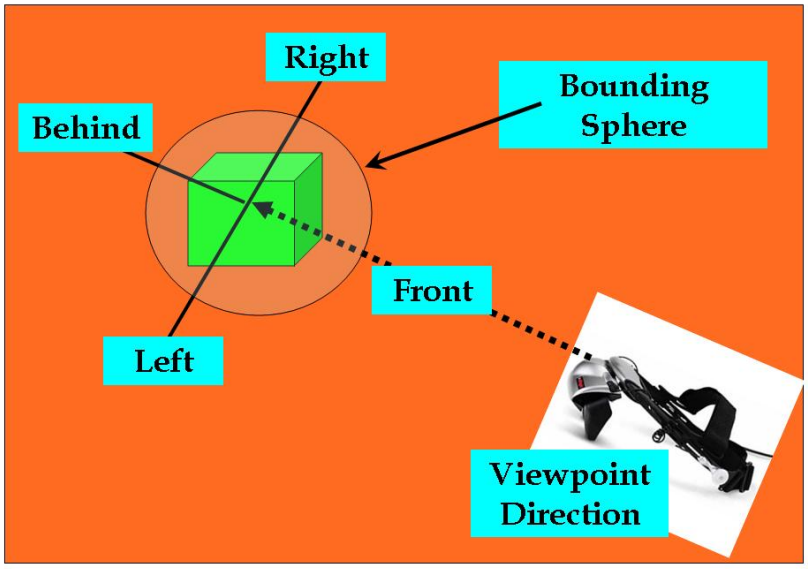


Figure 5.8 How the spatial predicates front, right, left and behind are defined.

Figure 5.9 shows how the software defines this point. Initially, a distance is calculated from the center of the object chosen. This distance is calculated by using the bounding sphere, or the smallest sphere that could be drawn encompassing the entire object. Since this distance would intersect with the outer surface of the object, an offset value, "SAFEDISTANCE" is added to the radius of the bounding sphere. This is "delta" and is used to place a point in the spatial reference indicated by the user. This delta value is used so that the selected point is offset form the object chosen to ensure that the robot does not collide with the selected object as it proceeds to the desired location.

The angle of the user viewpoint is calculated using the location of the fiducial marker the workspace is "attached" to. The variable "viewpos" is then calculated and used to find the camera position in one of four quadrants. The "findCameraPos" method returns an integer value indicating which of the four quadrants the user viewpoint is in. At this point, the system knows how the user is viewing the selected object and can then use this viewpoint to define the meaning of the spatial predicate used. In the current example, the system

```
osg::Vec3f toPlacePos, viewPos;
int cameraPos;
double delta, angle;

//  Find bounding sphere of object selected
osg::BoundingSphere bound = thisNode->getBound();

//  Delta is used to place spatial location external of the bounding sphere
delta = bound.radius() * SAFEDIST;

//  Find angle of viewpoint of user
osg::Matrix baseMatrix;
baseMatrix = baseTransform->getMatrix();
baseMatrix.invert(baseMatrix);
viewPos = baseMatrix.getTrans();

angle = atan2(-viewPos[0], -viewPos[1]);

//  Find camera position
//  Returns integer of camera position in one of four quadrants
cameraPos = Utils::findCameraPos(angle);

//  *action is spatial predicate used
//  in this case "behind"
switch (*action)
{
  case BEHIND :
    switch (cameraPos)
    {
      case 1 : toPlacePos[0] = delta * sin(angle);
               toPlacePos[1] = delta * cos(angle);
               toPlacePos[2] = MAZEZED;
               break;
      case 2 : toPlacePos[0] = delta * sin(angle);
               toPlacePos[1] = delta * cos(angle);
               toPlacePos[2] = MAZEZED;
               break;
      case 3 : toPlacePos[0] = delta * sin(angle);
               toPlacePos[1] = delta * cos(angle);
               toPlacePos[2] = MAZEZED;
               break;
      case 4 : toPlacePos[0] = delta * sin(angle);
               toPlacePos[1] = delta * cos(angle);
               toPlacePos[2] = MAZEZED;
               break;
    }
  break;

  // ...other cases
}

return toPlacePos;
```

Figure 5.9 Definition of position for "place via point behind this" command using the viewpoint of the user.

would place a point behind the selected object as defined by the viewpoint the user has of the selected object.

The ambiguity of the spatial command has now been resolved and a location behind the selected object identified. To reach common ground the robot gives a verbal response and the location is displayed to the user as an overlay in the AR environment. Therefore, the human is able to use natural speech and gesture and maintain situation awareness by receiving verbal and visual feedback of the robot's intended actions.

5.3 Summary

This chapter started by describing the components that make up the architectural design of the AR-HRC system. Each component was described and an example was provided to help explain how these components interact to provide the multimodal interaction required for robust human-robot collaboration.

The chapter concluded with a second example of deeper spatial dialog in an attempt to provide more detail of how the speech, gesture and viewpoint information is handled by the system. The overall design of the AR-HRC system is focused on providing an effective multimodal interface for robust human-robot collaboration. The examples presented are used to highlight the system functionality and operation using the software architecture described in Section 5.1.

Chapter 6

Multimodal Metric Study

In Chapter 3 it was shown that one aspect of an effective human-robot collaborative system is that it should support multimodal input and output. In that chapter, the idea of using AR technology to enable robust communication between the human and robotic system was presented. Chapter 4 discussed the development of a multimodal AR application and reported that this type of interaction in AR resulted in increased performance. Finally, Chapter 5 presented a system architecture for an AR interface for supporting human-robot collaboration.

The next step in the development of the AR-HRC system presented here was to determine what kind of speech and gestures would be best used to collaborate with a robot. Therefore, a Wizard of OZ (WOZ) study was conducted to enhance the development of robust multimodal interaction for the AR-HRC system. A description of a WOZ study is provided in Section 3.3.

The objective of the WOZ study was to find out what combination of speech and free hand natural gestures people would prefer to use when collaborating with a mobile robot on a navigation task. This chapter outlines the experimental design employed and the procedure followed in this study. Results are then presented and discussed.

6.1 Experimental Design

This section describes the environment in which the study took place and the tools used. It then presents the three user interface conditions examined. The

overall WOZ procedure employed is then described in detail. Finally, the study participants are presented.

6.1.1 Set Up

The experiment took place in two separate rooms. The two rooms shared a common wall with a one-way mirror that enabled the human wizard to observe the participants but did not allow the participants to see the wizard. The wizard station consisted of a Linux PC that ran the Gazebo 3D robotic simulation software from the Player/Stage project (Gerkey et al., 2003). Gazebo was used to provide the 3D simulated world of the robot and the robot itself.

The wizard used keyboard input in Gazebo to drive the simulated robot in correlation with the direction supplied by the users. The four arrows keys were mapped to forward and backward motion, as well as rotation in the left and right directions. This type of interaction was utilized so that the wizard could easily drive the robot with one hand. Each participant was tasked to guide the robot through the maze using one of three interface conditions. The three conditions used will be discussed in Section 6.1.2.

A windows PC was used to run a simple program that turned a set of keyed responses from the wizard into verbal responses. There was no GUI interface for the wizard, input to the program was by key selection and response was provided to the wizard through the command line window, and by the verbal response of the system. Using the Microsoft Speech SAPI 5 (MicrosoftSpeech, 2007) text-to-speech (TTS) functionality these responses were spoken by the system so the participants would believe that they were communicating with a robotic system and not a human, as required for this type of study.

The wizard also used these canned responses to alert the user when the experiment would begin, what modality would be used, if they had crashed and when they had finished the test. For example, if the robot crashed into a wall it would say "Ouch, I just ran into something." An attempt was made to use a bit of humor so that the human would feel as if the robot had a personality and thus would feel more comfortable in interacting with the robot as a collaborative member of the team. The wizard station can be seen in Figure 6.1.

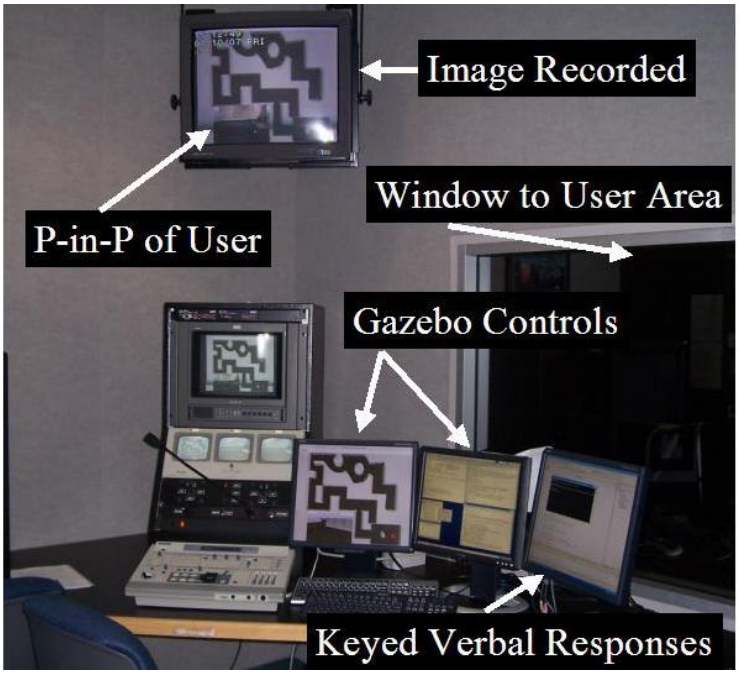


Figure 6.1 Wizard Command Center.

Video recording equipment captured both the video feed from the user's environment and the output from the Gazebo software. This recording consisted of the output from Gazebo in the background, the robot in its workspace, with an overlay of the user in the lower left corner in a picture-in-picture fashion. An example of the video captured for analysis is shown in Figure 6.2.

This recording also captured the speech used by the participants making it possible to correlate the speech and gestures used with the movement of the robot during analysis. This correlation enabled the tracking of the reference frame of the robot and which reference frame the participant was using. In this manner, it was possible to determine how the users interacted with the robot in terms of the reference frames used.

In the participant's environment, the video output from Gazebo was projected onto a screen that the user stood in front of. Two ceiling cameras were positioned so that the gestures used by the participant could be seen by the system and thus by the wizard. The wizard used this video feed to interpret the participant's gestures and drive the robot accordingly. The user's speech was picked up by a microphone in the ceiling. Speakers were placed near the user to provide the robotic voice output. The user environment is shown in Figure 6.3.

Figure 6.2 Example of video recorded during the WOZ study. Correlation with audio enabled analysis of reference frames used. The task was for the participant to guide a robot through the maze.

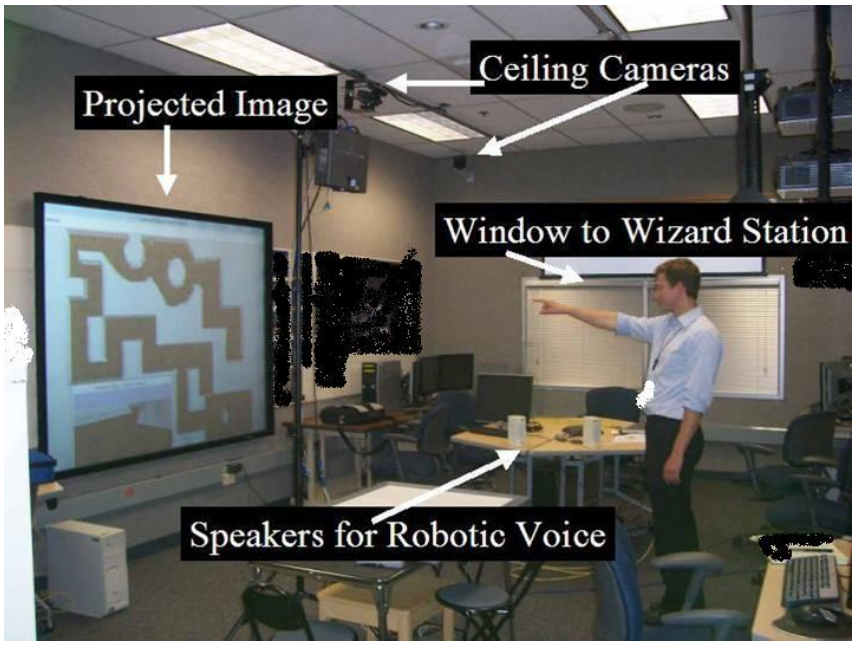


Figure 6.3 Environment for participants.

6.1.2 Experimental Conditions

In this experiment, three interface conditions were used:

- **Speech Only**: In this condition, the participants were allowed to use whatever speech input they wanted. However, no gesturing was used.
- **Gesture Only**: In this condition, the participants were allowed to use whatever gestures they wanted. However, no spoken dialog was understood.
- **Speech Combined with Gesture**: In this condition, the users were able to user whatever combined free hand gestures and spoken dialog they wanted.

6.1.3 Procedure

The study began by giving a demographic questionnaire to help evaluate how familiar the users were with robotics and speech interfaces, as well as determine age, gender, profession and educational experience. A pre-experiment questionnaire was then given to each participant. The objective of this questionnaire was to find out what type of speech and gestures the participants thought they would use prior to experiencing the system.

Users were asked to describe for each of the three experimental conditions how they would interact with a human collaborator. The questions given were modeled after those experienced on driving tests where the user is asked how they would complete a specified maneuver (LandTransportNZ, 2008). An example of such a question is shown in Figure 6.4.

The pre-experiment questionnaire used pictures instead of written questions. The reason pictures were used was to ensure that the participants would not be biased by the spatial language that would have been contained in written questions. In the pictured questions, the participant was to collaborate with the human to move from point A to point B, as shown in Figure 6.5.

The perspective in the pictures was varied to test what reference frames the participants would use. For example, if the picture depicted the reference

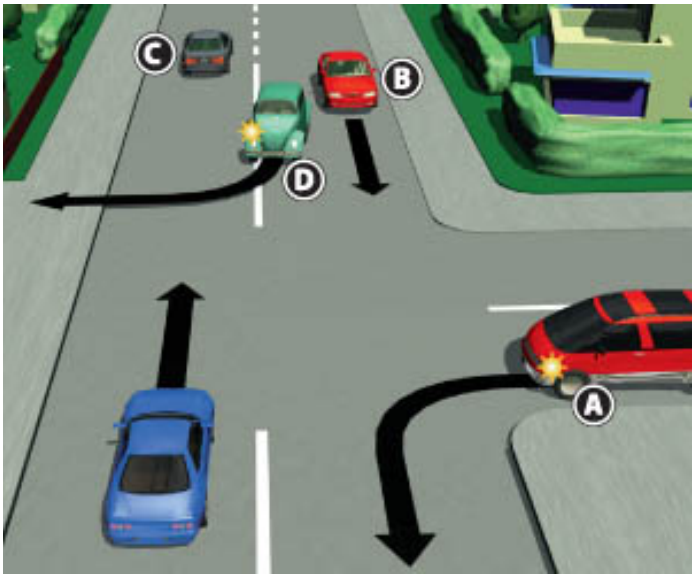


Figure 6.4 Example of a New Zealand driving test question that motivated the pre-experiment questionnaire design (LandTransportNZ, 2008).

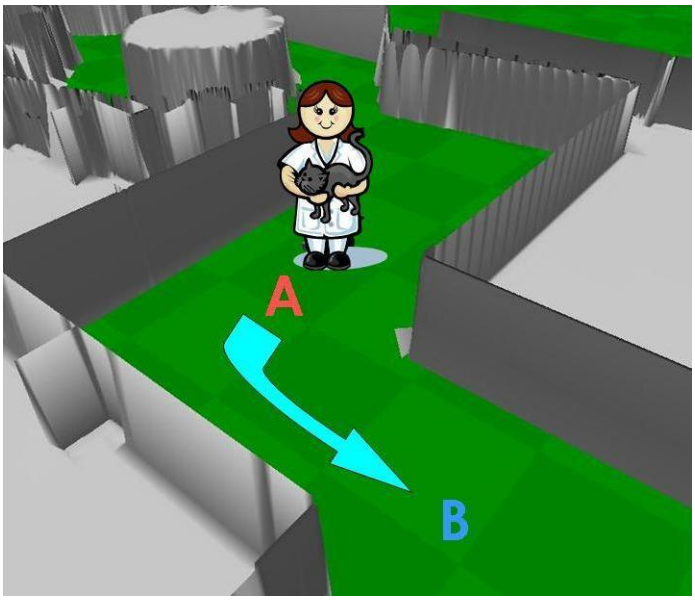


Figure 6.5 Example question to find out what speech, gestures, and speech combined with gestures participants would use as part of the pre-experiment questionnaire. In this example, the reference frame of the human pictured is purposefully not aligned with that of the participant.

frame of the robot aligned with that of the participant then it would be easy to predict that the user would refer to his or her own reference frame. However, if the reference frame of the robot was not aligned with the user then the objective was to see what spatial references would be used.

To determine if participants would communicate differently with a robot, as opposed to a human, a similar questionnaire was given out after the three trials for the different interface conditions were run. This post-trial questionnaire was similar to the pre-experiment questionnaire. However, instead of the human, the participants were questioned about how they would guide a robot from point A to point B.

The questions given for the three experimental conditions for both the human (pre-experiment) and robot (post-trials) randomly varied the reference frames used. However, there was one question that was given for each condition for both the human and robot cases. This question indicated the user had to have the robot or human go around an unidentifiable object or around a pizza. The latter case being something most participants could identify. Figure 6.6 shows the unidentifiable object and Figure 6.7 shows the identifiable object (pizza) that the user instructed the human and robot to go around.

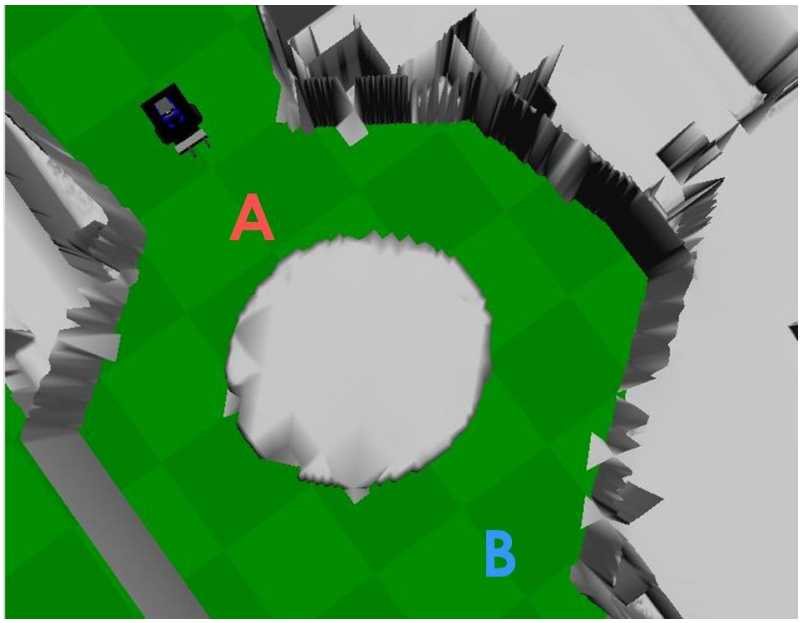


Figure 6.6 Question indicating robot to go around unidentifiable object.

The point of these questions was to determine what kind of language the participants would choose to use. For example, using "this" for the unidentifiable object or "pizza" for the object they could identify. In addition, participants were told that the human and robot had to stay on the green path and could not go into any gray areas. The questionnaires used in this study can be found in Appendix A.

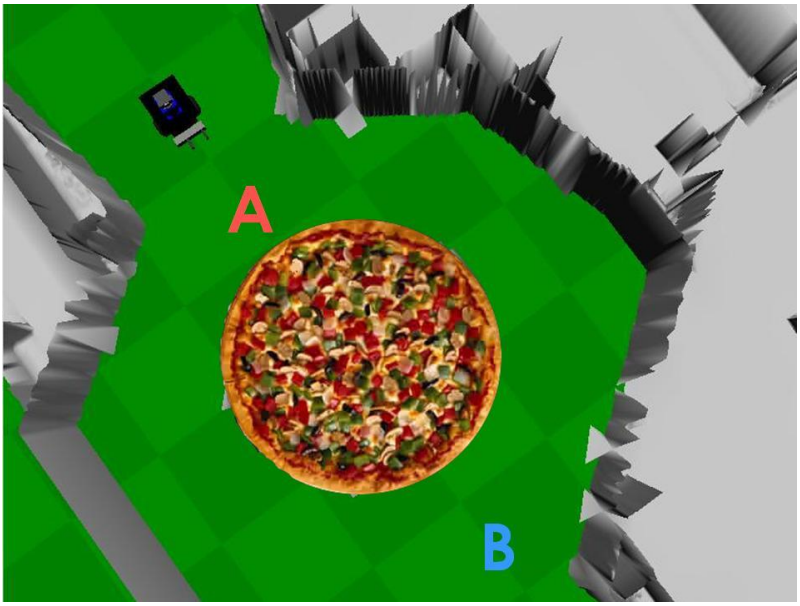


Figure 6.7 Question indicating robot to go around identifiable object.

After the pre-experiment questionnaire was completed, the task was explained to the participants. They were told that they would be working with a remote robotic lunar rover. The rover had experienced sensor failures and they were to collaborate with the robot to get it back to safety. This scenario was the cover story recommended by Dahlback et al. (1993), as discussed in more detail in Section 3.3.

Unknown to the users, the wizard was observing their speech and gestures and driving the robot accordingly. The same wizard was used for all participants to reduce the chance of varied interpretations of the participant's speech and gestures. The wizard responded to the user if speech or gestures were used that could not be understood with a statement such as "I'm sorry, I did not understand that." This response was presented verbally to the participants through the speakers in the user's environment. The wizard chose from a predefined list of responses when it was necessary to communicate with the participants.

A maze was created through which the user had to guide the robot. Participants were told that the robot was autonomous, but that its sensors had failed. For example, that it could not "see". The robot had an ego-centric camera that the human team member could see through but the robot itself could not make use of this camera. The participants had an exo-centric view

of the maze and robot in addition to the view from the camera mounted on the robot.

The objective for the participants was to work with the robot and guide it through the maze using one of the three interface conditions. Depending on the modality of the interface, users were told that the system was practically fluent in understanding spatial dialog, gesture or a combination of speech and gesture input. The participants were encouraged to use a wide variety of speech and gestures.

Each participant collaborated with the robot to get through the maze three separate times, one for each of the experimental conditions. The maze had multiple curves and forks so that the user would have to use a variety of spatial language. Similar to the pre-experiment questionnaire, the maze had unidentifiable objects in it that the robot had to maneuver around. The participants had both an exo (God-like) and ego (robot's) view of the workspace. The maze as projected for the user can be seen in Figure 6.8.

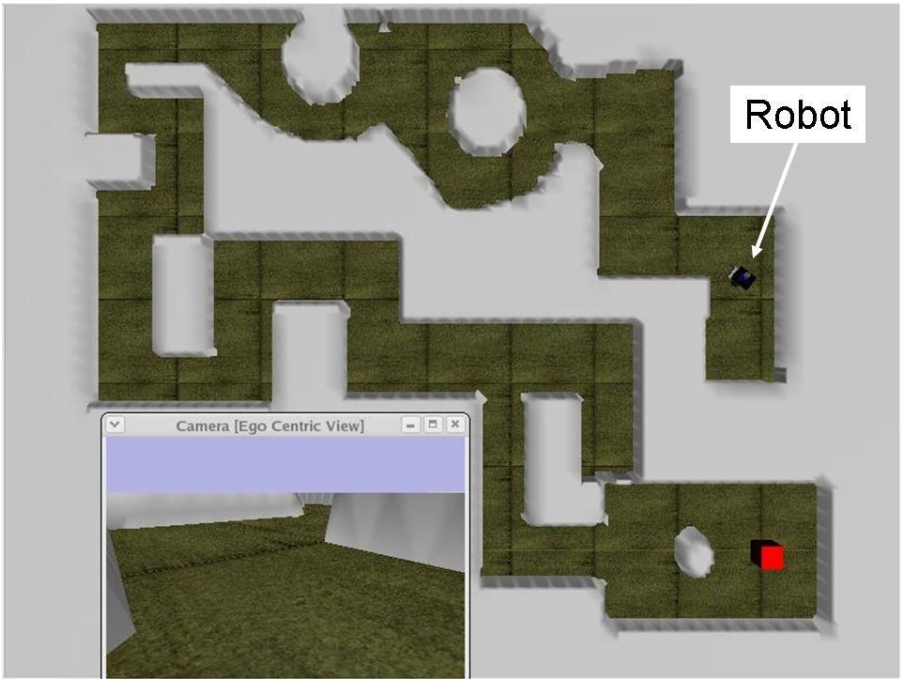


Figure 6.8 Example of maze for participants to guide robot through using speech and gestures. The robot was initially placed in the dead end (near where the robot is pictured) and the red box at lower right corner was the goal position. The lower left inset shows the robot's view.

The order of the conditions was counterbalanced between users to avoid sequencing affecting experimental results (Greenwald, 1976). A post-experiment questionnaire of 10 questions was given to the participants with answers provided on a Likert scale of 1 - 7. A Likert scale is one in which respondents indicate their level of agreement with statements that express a favorable or unfavorable attitude toward a concept being measured (Trochim, 2006). The questions were intended to gauge user satisfaction with the system and modality preference. All questionnaires are provided in full in Appendix A.

The sequence of events for the full user study was as follows:

- Demographic Questionnaire
- Pre-experiment Questionnaire with Human
- Explanation of Task
- Trial 1 (Robot)
- Trial 2 (Robot)
- Trial 3 (Robot)
- Questionnaire with Robot
- Post-Experiment Questionnaire

6.1.4 Participants

Ten participants were run through the experiment recruited from within the Lockheed Martin Space Systems Company, Advanced Technology Center, Sunnyvale, California, USA. The group consisted of nine engineers and one person from finance. There was one female and nine males all under the age of 25. The responses to the demographic questionnaire showed that overall the group was not familiar with robotic systems, not familiar with speech systems, and claimed they generally used gestures when speaking.

6.2 Results

The participants performed the task for each of the three modalities for a total of three trials per participant. Three objective measures were recorded, time

to completion, the number of crashes, and the distance the robot traveled. Although these three measures provide an indication of how the users performed, they were not the primary goal of this study. The goal of the study was to determine what kind of speech and gestures the participants decided to use. A secondary goal was to determine which interface the users preferred.

The following sections analyze the results from the experiment by first presenting the objective measures. Following the objective measures is an analysis of the pre-experiment questionnaire, where the participants indicated what types of interaction they thought they would use. An analysis of the speech and gestures used during the three conditions of the experiment is then provided. Finally, the results of the subjective questionnaires are given. All statistical analysis was performed using an analysis of variance (ANOVA) and post-hoc comparisons were done using Bonferroni correction (NIST, 2008), where warranted.

6.2.1 Objective Measures

The first objective measure considered is the time to completion. The multimodal condition had the shortest mean completion time of 428.5 seconds (Standard Error (SE) 41.14). However, an ANOVA test found there was no significant difference between conditions ($F_{2,27} = 2.17$, $p = 0.13$). Figure 6.9 shows the average completion times for the three conditions tested.

The next objective measure considered was the number of collisions. The multimodal condition had the lowest mean number of collisions with 5.5 (SE 0.22). However, an ANOVA test found there was no significant difference between conditions ($F_{2,27} = 1.88$, $p = 0.17$). Figure 6.10 shows the average number of collisions for the three conditions tested.

The final objective measure considered was the distance the robot traveled in completing the task. The three conditions resulted in similar distances traveled. An ANOVA test found there was no significant difference between conditions ($F_{2,27} = 0.27$, $p = 0.76$). Figure 6.11 shows the average distance the robot traveled for the three conditions tested.

These three measures were dependent on the user and not the modality of

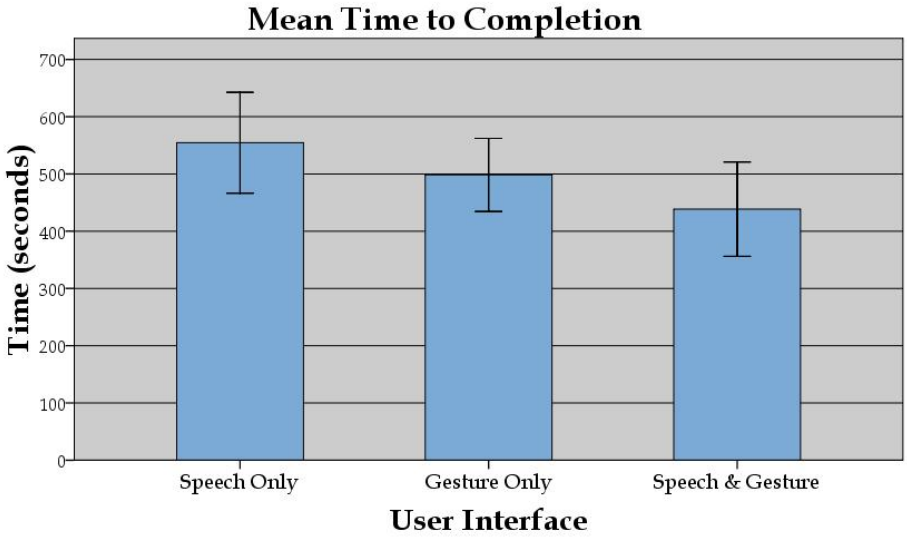


Figure 6.9 Average completion times for the three conditions.

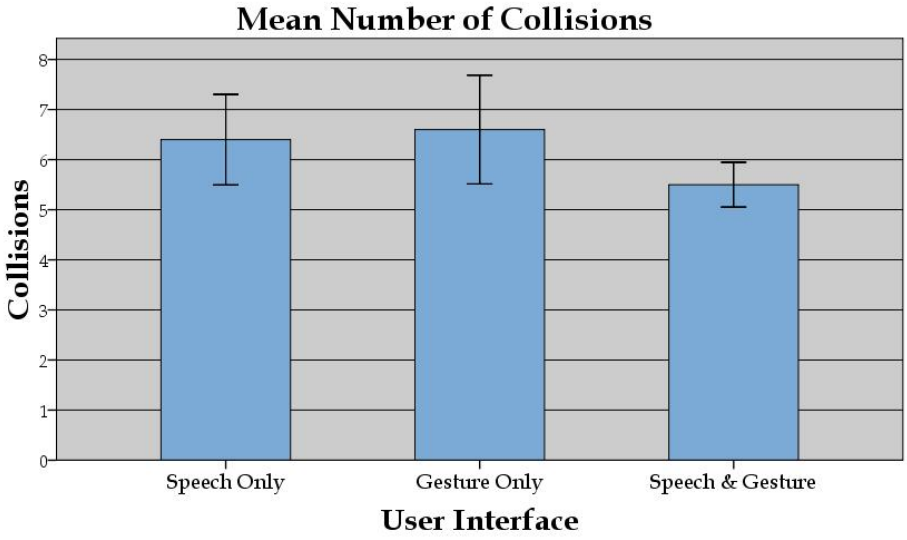


Figure 6.10 Average number of collisions for the three conditions.

communication. For example, if a participant crashed in one modality, then the user tended to crash in all three conditions. This result did not have an adverse effect on the study as the objective was not to determine which of the experimental conditions resulted in the best objective measures. The objective of the study was to find out what kind of speech and gesture participants would choose to use in a collaborative task with a mobile robot. These results are discussed in the following sections.

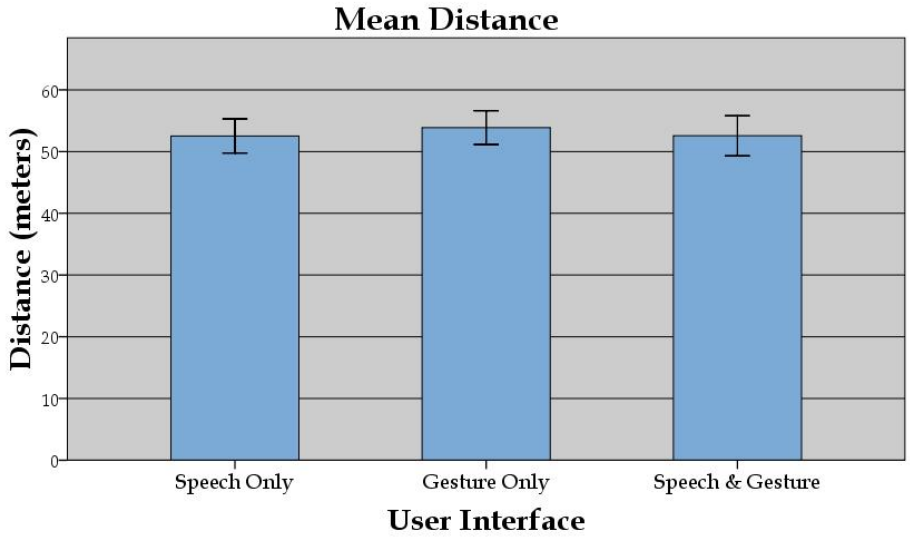


Figure 6.11 Average distance robot traveled for the three conditions.

6.2.2 Pre-Experiment Questionnaire

6.2.2.1 Speech Only Condition

When guiding the human from point A to point B for left and right turns (see Figure 6.5 for an example) users primarily used the term "turn" (9 right and 9 left), while one used "rotate" (right) and one used references to a clock, i.e. "7 o'clock" then "4 o'clock". Nine of the ten participants gave the directions in incremental steps, such as "walk, stop, turn, stop, walk, stop". Three participants included an angle with the command "turn", such as "turn right 90 degrees". To indicate forward movement users used a combination of the following commands: "move", "go", "forward", "straight" and "walk". Table 6.1 summarizes the speech commands used.

Action	Command Used	Modifiers
Forward	move, go, forward, straight, walk	none
Turn	turn, rotate	"30" degrees, clock directions
Stop	stop	none

Table 6.1 Speech commands used in pre-experiment questionnaire.

For the example of moving around the unidentifiable object and pizza, the participants used the same commands for both cases. This result was not anticipated. It was expected that the users would use go around "this" for the unidentifiable case, but no one did.

Eight participants gave incremental instructions, such as "forward, stop, turn right 45 degrees, stop, forward, turn left 45 degrees, stop, forward, turn left 45 degrees, stop, turn right 45 degrees, stop, forward, stop". Two participants used the preposition "around" and identified the pizza to go around. One user labeled the unidentifiable object as a pillar, while the instruction of the other user was "go around in a circle to your left". This use of the robots reference frame was consistent throughout the entire user study for all users.

6.2.2.2 Gesture Only Condition

Participants indicated they would use finger gestures (5 participants) or full arm gestures (4 participants) with the remaining user having a preference to use arm gestures analogous to those for riding a bike. Similar to the results for the speech only case, users gave directions in incremental steps. Right and left turns were instructed with either a full arm out in the appropriate direction or a similar instruction using only fingers. One participant indicated pointing to relative locations on a clock.

The gesture for stop was fairly consistent for all users. Hands up with palm out indicated stop. One participant used a quick up and down motion of the fingers to indicate stop. The final participant used a fist to indicate stop.

6.2.2.3 Speech and Gesture Condition

Participants combined the answers for the speech-only case and gesture-only case for the combined speech and gesture questions. Typically, the answers had the speech from the speech-only case complemented with the answers from the gesture-only case. This result is likely due to the users not wanting to repeat themselves. Hence, they used common answers that were already developed, instead of answering the questions from the scratch.

6.2.2.4 Comparison to Questionnaire with Robot

A similar questionnaire to the pre-experiment one was given out after the three trials were run where the human was replaced by the robot from the experiment. The intent of this questionnaire was to see how the users responses changed after running the experiment and to see if the communication with the robot differed greatly from that for the human figure. Indeed, the communication became more mechanized for the case with the robot. Each step was given incrementally with turns provided as discrete angles, except for one user who instructed the robot to "turn around the corner". Whereas in the case for the human a command was given to "go forward about 3 feet", the communication to the robot was simple, short and curt such as "move", "turn" and "stop" type utterances.

The gesture only case with the robot had only two participants using finger gestures as opposed to four for the case with the human. This result is most likely due to the fact that subtle finger gestures were less recognized during the course of the experiment. Additionally, users may have simplified, as with the voice modality, to ensure communication with the robot.

The speech combined with gesture case again was a combination of the other two modality responses. It is hypothesized that by this stage of the experiment the participants knew what questions were going to be asked and provided the quickest answers possible. This hypothesis is supported by responses such as "really, the same as before".

6.2.3 Experimental Results

6.2.3.1 Speech Only Condition

Participants tended to use the same verbal references for stop and turn as reported in the pre-experiment questionnaire. Stop was simply "stop" for turning they used "turn" and "rotate". A magnitude was sometimes associated with the turn and rotate commands, such as "turn right 30 degrees", while some of the participants followed a turn command with a stop command when the robot had reached an angle agreeable to the user.

A range of different commands was used to indicate the robot move forward, such as “walk”, “drive” and “inch forward”. The latter to indicate that the robot should move forward, but only a little bit. It was necessary at times for the users to have the robot move backwards. For this task, they used the two terms “backwards” and “reverse”.

One interesting result was the type of modifiers used. For example, to correct the robot when it had turned too far, users would say “back to the left”. If the robot had not rotated the amount the user expected, this was corrected with phrases such as “a little bit more”, “until I say stop” and “some more”. To quantify how far to travel phrases like “past that object” were used. In the test, the wizard responded to these types of ambiguous commands by moving the robot a small amount and waiting for a response from the user to determine if the robot had moved the amount intended.

Participants spoke in mechanized terms when they first started the experiment, as experienced by Perzanowski et al. (2003). If something unexpected happened, like a crash was impending, then the users would resort to communicating with the robot like it was a team member and not as if it were a robot. One explanation for this change was that when the users were conscious they were working with a robot they chose to speak to it in a manner they thought appropriate. This behaviour was seen in the questionnaires where the commands for the robot were more mechanized than for the human case. Once users felt comfortable with the system and its capabilities, they began to use more descriptive speech than just “turn”, “move” and “stop”.

Users commented after the experiment that “once I started using more complicated instructions than simple ‘go forward’ and ‘turn’ it became easier to control”. An example of this type of interaction was when one user kept the robot moving forward and would tell it to turn around the corners without stopping forward movement. Through the second half of the maze for this run, robot movement was much smoother, as opposed to the “turn, stop, move, stop” commands given in the first half of the maze.

Some participants used more descriptive phrases such as “go through the passageway in front of you” and “around the structure in front of you”. Participants also used “turn around the corner”, when the robot would stop after turning the corner, users would say “and keep going” to indicate the robot

should not stop after the turn. One user chose to speak to the robot by its given name, used "please" for each request and apologized to the robot when it crashed into a wall. These are indications that the user felt the robot was a true member of the team and spoke to it as if were a human team member. Table 6.2 provides a summary of the speech commands used during the speech only condition.

Action	**Command Used**	**Modifiers**
Forward	move, go, forward, straight, walk, drive, inch	past that, through the passageway in front of you, around the structure in front of you
Backward	backwards, reverse	none
Turn	turn, rotate	"30" degrees, back to the "left", a little more, some more, until I say stop
Stop	stop	none

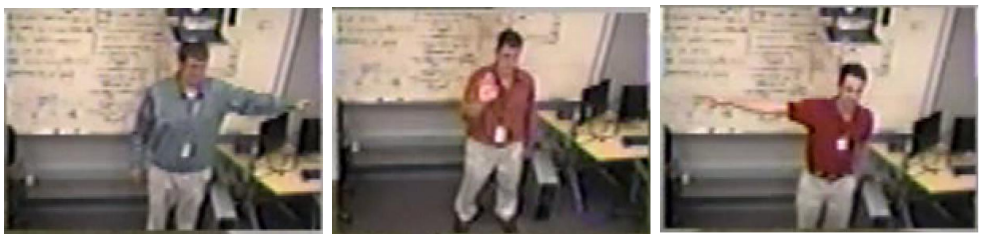
Table 6.2 Speech commands used in the Speech Only condition.

6.2.3.2 Gesture Only Condition

To have the robot move forward most users held their hand out at arms length in front of them. One user held the index finger up and then brought it down toward the screen in front of them to indicate the robot should move forward. Most users gave a gesture for the robot to move and then released the gesture. However, one user maintained gestures the entire time the move was desired. Thus, the entire time the robot was to move forward, this participant would keep his arm stretched out in front of him. Afterwards, perhaps as expected, the user commented on how tired his arms were at the end of the trial.

The gesture for stop was consistent between all users. Hands up, whether directly in front of the body or at full arms length, with palm towards the camera. This usage varied between users and also varied within the same trial of individual users. At times though, an impulsive stop was needed to prevent a crash and participants would have their palms pointing at the floor and wave them back and forth to indicate stop. Otherwise, the hand up palm forward method was the preferred method of gesturing stop.

Gestures for turning consisted of a full arm gesture to the side of the body that the user wanted the robot to turn in. All participants used the reference frame of the robot. Three users adjusted the degree of the turn by starting with the forward gesture (arm extended out in front of them) and defining the turn by how far their arm moved to one side. So a 45 degree turn would have the arm extended at full length and be in middle of having the arm extended out directly in front and completely to the side. Figure 6.12 shows three participants issuing a command for "left", "stop" and "right", where the images were extracted from the video captured as described in Section 6.1.1.



(a) Left (b) Stop (c) Right

Figure 6.12 Gesture commands for left (a), stop (b) and right (c).

One participant tried to use very slight hand gestures for turning. The wizard did not pick up on these slight gestures. Hence, the user interpreted this behaviour as the system needing exaggerated motion for gesture. Therefore, the participant turned their whole body in the direction of the desired turn. Another user at times would use only the index finger to indicate forward and turns, but full arm gestures were the majority.

6.2.3.3 Speech and Gesture Condition

Users tended to use the same methodology for guiding the robot in the multimodal condition as in the speech only and gesture only conditions. This methodology for seven participants consisted of combining the techniques used in the verbal only and gesture only trials to guide the robot, but doing so in a sequence of incremental steps "go, stop, turn, stop, go" etc. Three participants used more complex communication, such as "go around this" while using a full arm gesture to indicate a turn, or "go around the corner to your right", again while gesturing using a full arm extended to the side indicating to turn.

The result of this type of communication was more fluid motion of the robot. When more descriptive communication was used, there were fewer stops for the robot which resulted in decreased time to complete the task. The three participants who used the more descriptive communication that resulted in fewer stops all had completion times far less than the average. The average completion time for the multimodal case was 438.5 seconds. Tellingly, the three users with fluid robot motion had significantly lower completion times of 272, 291 and 298 seconds. This result shows that using more complex communication enabled fluid robot motion that decreased completion times due to enhanced communication and collaboration.

6.2.4 Post-Experiment Questionnaire

The post-experiment questionnaire consisted of ten questions to which the participants responded using a 7 point Likert scale. An answer of 1 indicated an answer of "very much so" to the statement and an answer of 7 indicated an answer of "not at all". The first seven questions are discussed next, followed by an analysis of the last three questions, which asked the users if they considered each of the three interface conditions as the best interface.

- Q1 *Do you feel the system reacted the way you thought it would before you began the experiment?* The mean response for this question was 3.40 (SE 0.62). A two tailed t-test showed that there was no significant difference from the mid point value of 4, with $p = 0.36$. Therefore, although the mean of the participants answers indicated that the participants felt the system reacted as expected, the result was not significantly different from a neutral response.

- Q2 *How well did you feel the system understood your verbal spatial references?* The mean response for this question was 1.60 (SE 0.16). A two tailed t-test showed that there was significant difference from the mid point value of 4, with $p < 0.05$, indicating that the participants felt strongly that the system did understand the spatial language that they used.

- Q3 *How well did you feel the system understood the gestures you used?* The mean response for this question was 3.30 (SE 0.58). A two tailed

t-test showed that there was no significant difference from the mid point value of 4, with p = 0.26. Therefore, although the mean of the participants answers indicated that the participants felt the system understood their gesturing, the result was not significantly different from a neutral response.

- Q4 *How well did you feel the system reacted the way you wanted it to?* The mean response for this question was 2.90 (SE 0.48). A two tailed t-test showed that there was significant difference from the mid point value of 4, with $p < 0.05$, indicating that the participants felt the reaction of the system was what they had intended it to be.

- Q5 *Do you feel the use of gestures helped you communicate spatially with the system?* The mean response for this question was 3.40 (SE 0.50) . A two tailed t-test showed that there was no significant difference from the mid point value of 4, with p = 0.26. Therefore, the result was not significantly different from a neutral response.

- Q6 *Did you have confidence speaking to the system?* The mean response for this question was 2.20 (SE 0.59). A two tailed t-test showed that there was significant difference from the mid point value of 4, with $p < 0.05$, indicating that the participants felt strongly that the system responded well to verbal input.

- Q7 *Did you have confidence gesturing to the system?* The mean response for this question was 3.50 (SE 0.59). A two tailed t-test showed that there was no significant difference from the mid point value of 4, with p = 0.43. Therefore, although the mean of the participants answers indicated that the participants felt that the system responded well to gesture input, the result was not significantly different from a neutral response.

The results of the post experiment questionnaire can be seen in Table 6.3. These results show that participants felt the system understood verbal spatial references very well and that the system reacted the way they expected it to. Users also felt comfortable speaking to the system. Responses for how well the users felt the system understood gestures were neutral. This result should be expected since the wizard was able to fully understand the speech used, but had to interpret the gestures, which took time and, when gestures were ambiguous, the wizard was not always able to understand them.

	Mean	**Std Error**	**T-Test (p)**
Q1: Do you feel the system reacted the way you thought it would before you began the experiment?	3.40	0.62	0.36
Q2: How well did you feel the system understood your verbal spatial references?	1.60	0.16	< 0.05
Q3: How well did you feel the system understood the gestures you used?	3.30	0.58	0.26
Q4: How well did you feel the system reacted the way you wanted it to?	2.90	0.48	< 0.05
Q5: Do you feel the use of gestures helped you communicate spatially with the system?	3.40	0.50	0.26
Q6: Did you have confidence speaking to the system?	2.20	0.59	< 0.05
Q7: Did you have confidence gesturing to the system?	3.50	0.59	0.43

Table 6.3 Summary of questionnaire responses. Questions were posed on a seven point Likert scale between 1 = Very Much So and 7 = Not at All.

The final three questions asked the participants which of the interfaces they felt was the best, with responses given on a Likert scale from 1 (very much so) to 7 (not at all). The results are shown in Figure 6.13. An ANOVA test resulted in ($F_{2,27} = 4.09$, $p < 0.05$) showing there was a significant effect due to condition. Pairwise comparison using Bonferroni correction ($p < 0.05$) revealed significant differences between the multimodal (Speech and Gesture) and the Gesture Only conditions. However, there was no significant difference between the Speech and other two conditions. Therefore, although the mean of the participants answers indicated that the participants felt the Speech and Gesture condition was the best, the result was not significantly significant.

6.3 Discussion

The primary objective of this study was to determine what kind of speech and gestures people would use to interact with a mobile robot. For this reason, participants were encouraged to try a variety of spatial references and gesture interactions. Users were encouraged not to repeatedly use the same interaction technique once they found a given technique worked well for them, but to try new techniques instead to see if the system would understand them.

Given the opportunity, participants used natural speech and gestures to work with a robotic team member. Initially, with no instructions given on what

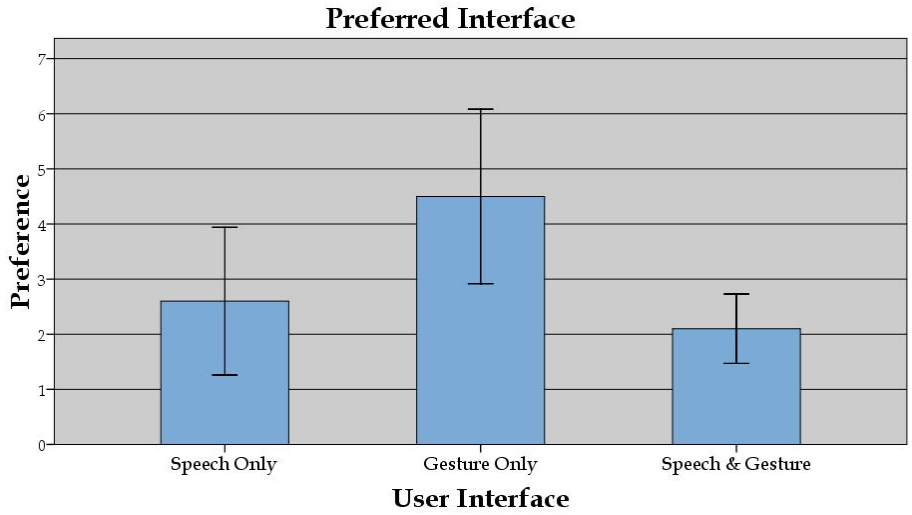


Figure 6.13 User modality preference. Questions posed to determine if the participants felt each interface was the best on a seven point Likert scale between 1 = Very Much So and 7 = Not At All.

type of speech and gestures to use, the participants communicated with the robot in a manner they thought the robot would understand. This manner was short, mechanized terminology, such as "rotate", "stop", "forward", "stop". However, once the participants learned they could communicate in a natural fashion they did so, "go around that corner in front of you", and commented on the natural and intuitive nature of the interface.

Users preferred full arm gestures to indicate forward and turning motions. To prevent users from getting tired of making these types of exaggerated gestures, it is important that the system react in such a way so that the user does not have to maintain the gesture. For example, to turn right a user should be able to point to their right using a full arm gesture and then return to a normal relaxed pose. The system should react by initiating a turn and continuing to do so until a command is received to stop. One comment was made that the user preferred speech because then their arms "would not get tired", so it is important to think about ergonomics when designing gestures into a system.

A gesture for turning should also define the magnitude of the turn. Three participants used this type of gesturing and two others commented it should be incorporated. Two of the three that actually used this type of gesturing held one arm out for move forward, then used that arm to start moving to one side to indicate to begin to turn to at what degree. The other participant

used one arm forward to indicate move forward and then used the other arm to continually make turns. When the turn would go from right to left, the user would change which arm was used for the forward motion (always maintaining this forward motion) and use the appropriate arm for gesturing a turn and its magnitude. This type of interaction is shown in Figure 6.14, this image was also extracted from the recording described in Section 6.1.1.

Figure 6.14 Participant simultaneously gesturing forward and to the right resulting in fluid robot motion around corners.

The participants commented that the verbal responses from the robot were helpful as it let them know what was happening and what was going to happen. However, users did comment that they would have liked to have had the robot tell them when a collision was imminent. It should be noted that in this study the sensors from the robot were supposed to have failed. Thus, it would have not been practical to have the robot tell the users a collision was imminent, although it would have been helpful for the user.

Participants were not consistent with the speech and gestures they used. During a trial one user would use various forms of speech and gestures to mean the same thing. For example, they would hold one hand up palm out for stop, but then also hold arms out with palms toward the floor waving their hands back and forth. The impact of this result is that for a system to be robust, it must be able to understand not only the various forms of communication between different users, but be able to adapt to or understand the changing communication of each individual user.

One participant commented that it would have been nice to interact with the visual representation of the robot‘s environment. The user would have

liked to have been able to touch a point on the screen and tell the robot to go "there". This comment was encouraging news for this research as that is exactly the kind of interface envisioned, using Augmented Reality as a means for enabling a user to pick out a point in the 3D representation of the robot in its work environment and referring to it as "here" or "there".

An important component of a WOZ study is that the participants must believe the system is fully functional. The users cannot know that a wizard is really running the system. Post experiment discussions revealed that all ten participants thought that they were interacting with a real functioning system and were not aware of, nor suspected, that a human was involved in the running of the experiment.

6.4 Design Guidelines

The results of the study discussed in this chapter provide a few design guidelines for the AR-HRC system:

- The human dialog should be flexible and adaptable to various human users.
- The system should provide feedback to the user indicating how the multimodal interaction was interpreted.
- The system should allow for interaction with the virtual representation of the robot's world.
- High level communication should be incorporated since it results in smoother robot motion.

6.5 Summary

This chapter described the experimental design of a Wizard of Oz (WOZ) study conducted for HRI. The results from the pre-experiment questionnaire were presented. Experimental results were then discussed and the responses to the post experiment questionnaire were analyzed. A discussion of these

various results and the impact they had on the design of the AR-HRC system ended the chapter.

The primary objective of the study discussed in this chapter was to find out what kind of speech and gestures people would use to interact with a mobile robot. Given the opportunity, participants used natural speech and gestures to work with a robotic team member. Initially, with no instructions given on what type of speech and gestures to use, the participants communicated with the robot in a manner they thought the robot would understand. This manner was short, mechanized terminology. However, once the participants learned they could communicate in a natural fashion they did so and commented on the natural and intuitive nature of the interface.

It was observed that when participants used more descriptive communication behaviour, the result was fluid robot motion and reduced completion times. Participants also commented on the usefulness of having the robot verbally respond to enable them to maintain awareness of what the robot was doing and what it was "thinking". Therefore, a multimodal approach to human-robot communication results in the most effective communication taking place, thus enhancing the collaborative interaction.

Chapter 7

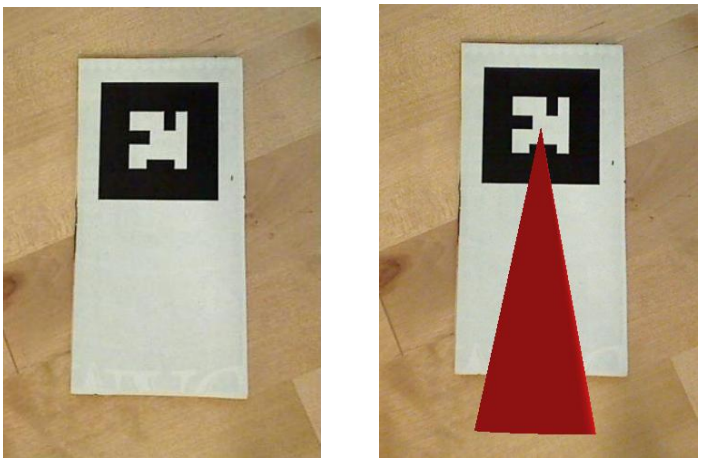
Integration with a Mobile Robot

This chapter outlines the integration of a mobile robot into the AR-HRC system. The interaction techniques are outlined first, and then the integration with a mobile robot is described. Finally, an example is provided of how a user would collaborate with the mobile robot subsequent to experimental evaluation in a later chapter.

7.1 Interaction Techniques

In this research, gesture interaction involving the use of the real world paddle combined with speech input is used in a multimodal interface. The paddle can be used as a pointer, enabling the human to point into the 3D virtual world of the robot and select a point or object. A second modality enables the human to use the paddle for natural gestures. The user is able to issue a verbal command to switch between these two modalities of paddle gesture interaction.

The real world paddle is flat and has a fiducial marker on both sides which enables the vision system to see the marker no matter which way the user holds the paddle. The paddle shown in Figure 7.1 is in pointer mode. The red cone is the virtual representation of the paddle in the AR environment. This cone is "attached" to the real world paddle through the use of the ARToolKit tracking library (ARToolKit, 2008). Thus, when the user manipulates the real world paddle, the movements are mapped one to one to the movements of the virtual cone.



(a) Real Paddle (b) Virtual Pointer

Figure 7.1 The real world paddle (a) and the virtual pointer attached to the paddle in the AR environment (b).

The virtual pointer is the visual cue used by the human to select locations and objects in the virtual world. When the virtual pointer intersects other virtual content in the scene it is occluded, providing the user with the perceptual cues necessary to determine precisely the point or object selected by the virtual pointer. In this mode, the human is capable of reaching into the virtual representation of the robot's world to select objects and points in space.

A second modality of gesture interaction enables the human to use the paddle for natural gesture interaction. The definition and development of this gesture interaction was informed from the results of the WOZ study discussed in Chapter 6. Natural gestures have been defined to communicate to the robot to move forward in a straight line, turn in place with no forward motion, move forward while turning either left or right, back up and stop. At any time, the user can issue a verbal command for these motions resulting in a true multimodal experience.

The system determines the orientation of the paddle relative to the user's point of view and uses this information to define the gesture. For example, if the paddle is held straight out in front of the user and the orientation angles of the paddle fall within certain defined threshold values, then the system interprets this gesture interaction as a move straight forward command. Similarly, when the paddle is moved to either side of straight in front of the user the

system calculates the angle from straight ahead and converts this information into a turn.

To turn the robot in place the user starts from the straight up position and rotates their arm about their elbow to the right or left. To go in the reverse direction the user places the paddle in a straight up position. Any position of the paddle not specifically defined is interpreted as a stop command and is relayed to the user by displaying a stop sign.

When the paddle is used for natural gestures the virtual pointer does not appear. The system keeps the user informed of what paddle gesture is active by displaying the appropriate icon on the paddle. When the user switches the paddle mode from pointer to gesture, the red virtual pointer is replaced by one of these icons. Figure 7.2 shows the icons displayed for the various paddle-gesture commands.

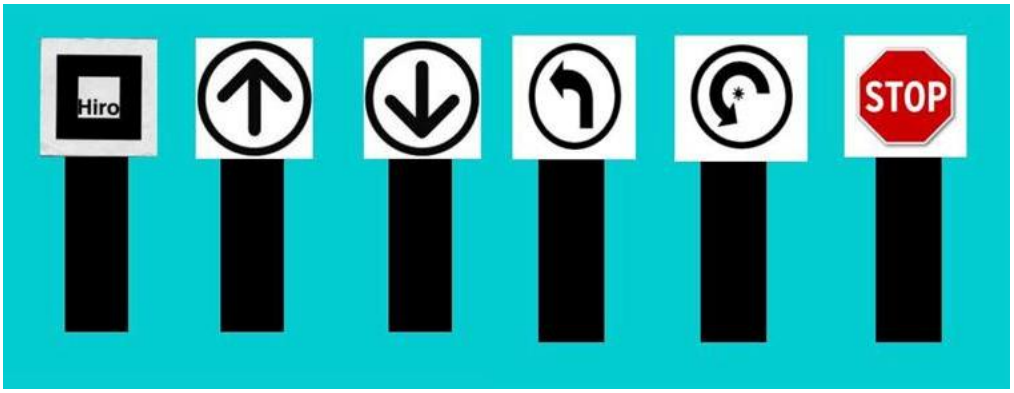


Figure 7.2 The image on the far left is of the paddle as seen in the real world. The remainder of images show the AR view showing, from left to right, forward, reverse, forward turn left, turn left in place and stop.

The viewpoint direction of the user is determined through the computer vision techniques made available from the use of the ARToolKit libraries. The line of sight of the user into the virtual world is computed by calculating the position of the camera mounted on the HMD, this calculation represents the user's viewpoint direction. By comparing this position to the position and orientation of the marker set that represents the robots virtual world the user viewpoint direction can be determined.

The viewpoint direction of the user in the AR environment is then used to define spatial references such as "behind" and "to the right of" objects

selected using the real world paddle in the pointer mode. By knowing where the user's viewpoint is in reference to the objects in the virtual scene these spatial references can be defined in the reference frame of the user. This information is then converted into the reference frame of the robot. The conversion is made possible through the use of AR, which provides a common reference frame for both the robot and human collaborators. The desired location is then sent to the robot where it uses its autonomous capabilities to move to the position in the real world.

Another benefit of using AR as a means to mediate the communication between the robot and human is the ability to smoothly transition from an exocentric (god's eye view) to an egocentric view. This means the user can smoothly transition from a bird's eye view of the robot in its environment to the view provided by the robot's camera, and vice-versa. The user is able to issue a verbal command to switch from one viewpoint to the other. This ability to view the robot's world from two vantage points increases the situation awareness of the human by allowing the scene to be viewed from the various vantage points.

The human and robot are able to create a path plan and review this plan before the robot is set in motion. The human is able to point to a location in 3D space and issue a command such as "go here". Alternatively, the human can select an object with the paddle and instruct the robot to "go behind this". The robot then displays its path trajectory in the AR environment to reach common ground with the human by showing its intended actions. Thus, the human is immediately able to determine if the robot understood the intended plan.

The plan under development can easily be modified through the use of via points. Additional nodes in the path can be added or deleted through the use of spoken dialog and gesture interaction. These way points help to ensure smooth motion and obstacle avoidance. Each time the path is modified, the robot displays its updated path as overlays in the AR environment and verbally acknowledges that the path has been modified.

These visual and verbal responses were incorporated into the AR-HRC system as a result of the user study discussed in Chapter 6. The verbal acknowledgments are randomly selected from a pool of appropriate responses so

as not to tire the user with the utterance of the same verbal response. In addition, the user can choose to show the planned path or hide the trajectory if the overlay interferes with viewing other important parts of the environment.

An important aspect of the interaction dynamics of the system is that the user can ask the robot to review the plan in its entirety prior to having the robot execute the plan. This review consists of the virtual representation of the robot running through the plan, enabling the user to see what is happening through the AR overlays. This step enables the user to identify possible problems with the plan. During review, if the robot determines a crash is possible, it stops reviewing the plan and asks the human if it is safe to continue with plan review. In this manner, the robot and human are able to work together in a collaborative manner to create a plan that they both "think" is appropriate.

The ability to review the plan prior to execution provides the means for detection of unexpected situations and the identification of probable collisions. The result is that the motion of the robot, once it executes its collaboratively designed plan, is smoother and collision free. This type of interaction between the robot and human in creating the path plan through shared verbal and spatial references, identifying possible problems with the plan and eliminating these issues, highlights the collaborative nature of the human-robot interface. Results from the study discussed in Chapter 6 showed that high level communication of this nature resulted in smoother robot motion and more natural interaction for the human.

During execution of the plan the robot acts autonomously. The human is able to monitor the progress of the plan through the AR environment as the robot updates the system with its internal state and position information. The human at any time can interrupt plan execution and modify the plan if the situation warrants it. Similarly, the robot is able to stop execution and ask the human for help if a situation arises that it cannot resolve on its own. At any point during planning or execution, the human has the ability to abort the plan. In this manner, the level of autonomy of the robotic system is effectively varied depending on the given situation.

It's important that the human is kept abreast of the internal state of the robotic system, as well as the state of the system as a whole. This knowledge increases situation awareness by helping the human to realize how the robot's

action will affect its environment and its ability to complete a requested task. The human is kept aware of the internal state of the robot in two ways. First, the robot verbally alerts the user if its internal state might put the completion of the task at risk. For example, if the robot's battery level sinks to a dangerous level, then the robot will verbally alert the user.

The user is also kept aware of the internal state of the robot on a more constant basis through the use of a heads up display (HUD). The HUD is part of the graphical overlay of the real world view of the user. In this display, the user is presented with internal state parameter information such as battery level, sensor readings, motor speeds and communication status. The state of the paddle modality is also presented for the user. Figure 7.3 shows an example of the information displayed in the HUD.

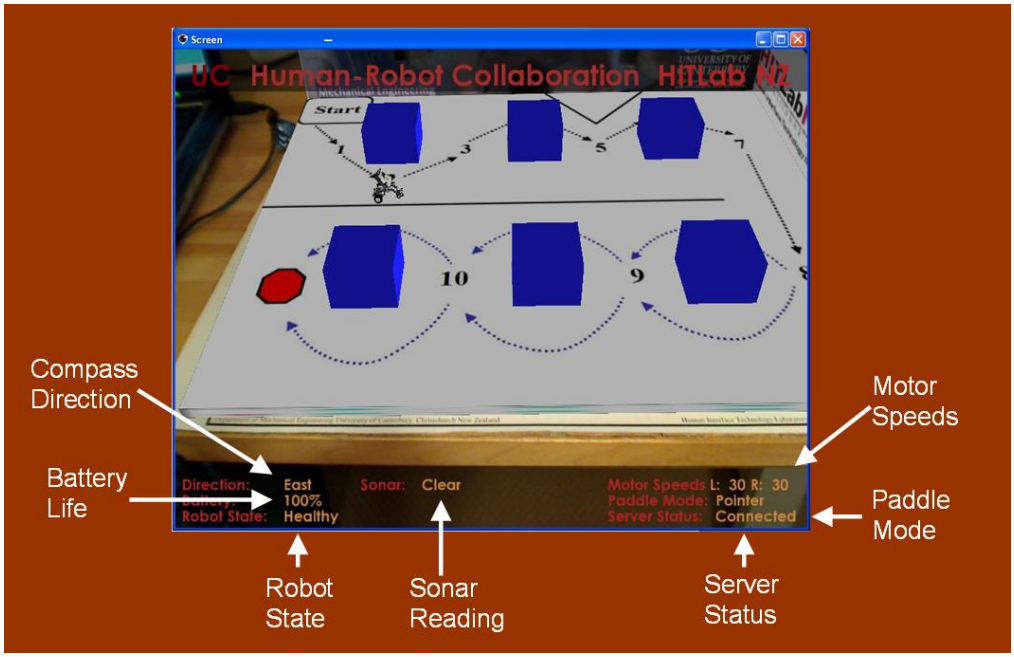


Figure 7.3 Heads Up Display (HUD) presented to the user as a graphic overlay to heighten situation awareness of robot and system state.

7.2 Integration

As a case study, a Lego MINDSTORMS^{TM} NXT (TheLegoGroup, 2007) mobile robot in the Tribot configuration was used as a mobile robot for collaboration. The NXT robot can be seen in Figure 7.4. To incorporate the mobile robot into

the system the NXT++ libraries (NXT++, 2007) were used. These libraries represent an interface to the MINDSTORMS^{TM} robot written in C++ that enables a PC to communicate with the robot through a Bluetooth connection. A Lego MINDSTORMS^{TM} robot was chosen because it is a simple low cost platform to prove out the functionality of the AR-HRC system.

Figure 7.4 The Lego MINDSTORMS^{TM} NXT robot (TheLegoGroup, 2007) used for integration into the AR-HRC system.

The configuration of the NXT robot used had one ultrasonic sensor on the front to sense objects and measure the distance to them. The robot also had a touch sensor on the front that would stop the robot if triggered to avoid colliding into objects. The limited sensing ability of the robot allowed the use of spoken dialog to increase collaboration in ensuring the robot took a safe path.

The AR-HRC needs as an input from the robotic system the location of the robot relative to its environment. The NXT robot sends this information as the motor counts to the AR-HRC system through the Bluetooth connection. The robot is zeroed out at the initial position and placed at a known location in the real world environment. From this location, any motion was calculated through the change in motor counts. However, using motor counts leads to rather large errors when the robot moves large distances. For this integration, these accumulated errors did not inhibit the interaction of the collaborative environment, which was the main focus. However, for the evaluation of the system in Chapter 8 a simulated robot was used so that the evaluation of

the system would be on the interaction with the robotic system and not the peculiarities of the NXT robotic platform.

An example of using dialog to ensure safe robot motion would be when the robot had to back up. With no rear sensors the robot was unable to determine if a collision was imminent. In this case, the robot asked the human if it was ok to move in reverse without hitting objects in its environment prior to commencing movement. Once the robot received confirmation that the path in the reverse direction was clear, it began to move in the reverse direction. Since the robot had to ask for guidance to complete the reverse maneuver, the user was aware that the robot might need assistance. It was then assured that the user has maintained spatial awareness, which in turn enabled a collaborative human-robot exchange and the resulting safe execution of robot motion.

7.3 Interaction Scenario

In this section, a scenario where the human interacts with the NXT robot using the AR-HRC system is described. The human interacts with the AR-HRC system at a command center remotely located from the robot. The user wears a HMD with a web cam attached to it, as well as a noise canceling microphone. These items are connected to a PC running the AR-HRC software.

A fiducial grid is set out on a table that serves as the position where the robots virtual world will be displayed to the user through the user of graphic overlays of the web cam input. The user interacts with the system using speech and gesture interaction with a real world paddle that also contains fiducial markers for AR tracking. A user in such a setting is shown in Figure 7.5.

In the HMD, the user sees overlaid on top of the real world view a 3D graphic of the robot in its environment. The session begins with the robot signaling to the human it is ready to collaborate.

Robot: *Good day. What would you like to do today?*

Human: *Let's make a plan.*

Robot: *Ok, let's start with the first point then.*

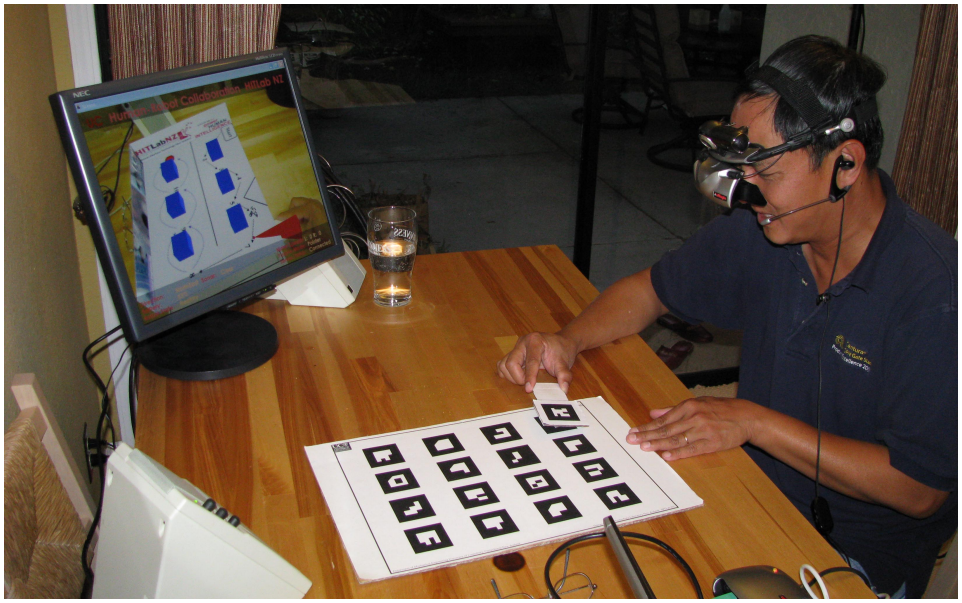


Figure 7.5 User interacting with the NXT robot using the AR-HRC system.

The human gestures into the AR environment with the real world paddle and selects a point where the robot is to go.

Human: *Please go here.*

The AR-HRC system seeks to define the ambiguous term "here" by first checking if the fiducial marker for both the robot's environment and the paddle are in the user's field of vision. If both are found, the 3D location of the paddle is then calculated. This 3D location is then translated into the reference frame of the robot. The point selected is displayed as a sphere and the straight line trajectory from the robots current position to the selected point is overlaid on the AR world environment. This interaction is shown in Figure 7.6.

Robot: *I'm showing the point selected and displaying our plan.*

The human is immediately able to see the intentions of the robot and reach common ground. The human then progresses with the plan by adding points to the trajectory for the robot to follow. The human selects one of the objects in the robots environment, the selected object turns a different color to let the user know which object has been selected.

Human: *Place via point in front of this*

Figure 7.6 User selecting location in remote robot environment. User verbally requests robot to proceed to point selected.

The AR-HRC system determines the human viewpoint direction by calculating the orientation of the fiducial marker representing the robot's work environment. This orientation tells the system where the camera is in relation to the fiducial marker. Once the viewpoint direction of the user is known, the system then calculates where the next point should be placed based on the spatial reference used, "in front of" for this example. A sphere is placed at this point so the user can see where the robot will go. The path trajectory is then updated to include this additional point. The result of this interaction is shown in Figure 7.7.

Robot: *Our plan now goes in front of the selected object.*

If the human is not satisfied with the placement of this point, then it can be deleted from the plan.

Human: *Please delete last via point.*

The point last created is deleted, the plan trajectory is updated and is displayed for the user to see.

Robot: *Deleted last via point.*

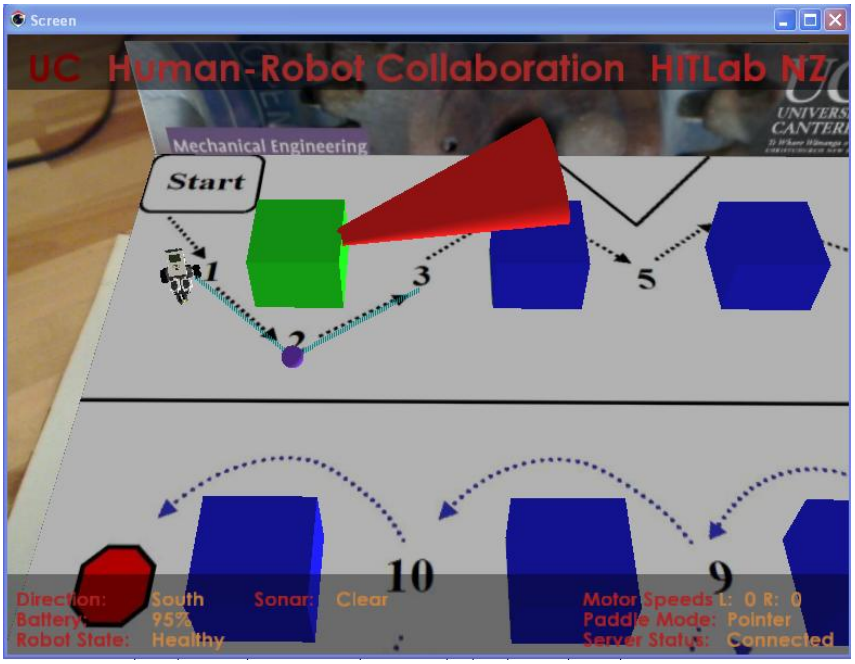


Figure 7.7 User selects virtual object with real world paddle. User verbally requests the robot to add a via point at the spatial location in front of the object. The AR system updates the trajectory enabling human to immediately reach common ground.

The human is able to continue interactively planning with the robot using spatial dialog and paddle gestures. Once a plan has been defined, the human is able to ask the robot to review the plan prior to sending the plan off for execution.

Human: *Let's review the plan.*

Robot: *Ok, reviewing plan.*

The robot now runs through the path as planned. This review is displayed in the AR environment allowing the human to monitor how the plan will be played out. At any point, the human can interrupt the review and modify the plan if warranted. If the plan plays out to the satisfaction of the human, then the human can send it off to the live robot for execution.

Human: *Please execute*

The plan is then sent to the live robot for execution. As the robot executes it updates the system with its internal state and location information. The graphics in the AR environment are updated to allow the human to watch the robot as it completes the plan. If the robot runs into an unexpected situation it can ask the human for help.

Robot: *I sense something in the way, can I continue?*

The human is able to determine how close the robot is to objects in its environment. If needed, the human can move the fiducial grid that the robot's environment is "attached" to to get a better perspective. Figure 7.5 shows a user who has rotated the fiducial grid to gain a better perspective of the robot's work environment. If the human determines the robot can continue, then the human instructs the robot to do so. Figure 7.8 shows the robot stopped near an object. The human is able to determine whether the robot should continue on its path.

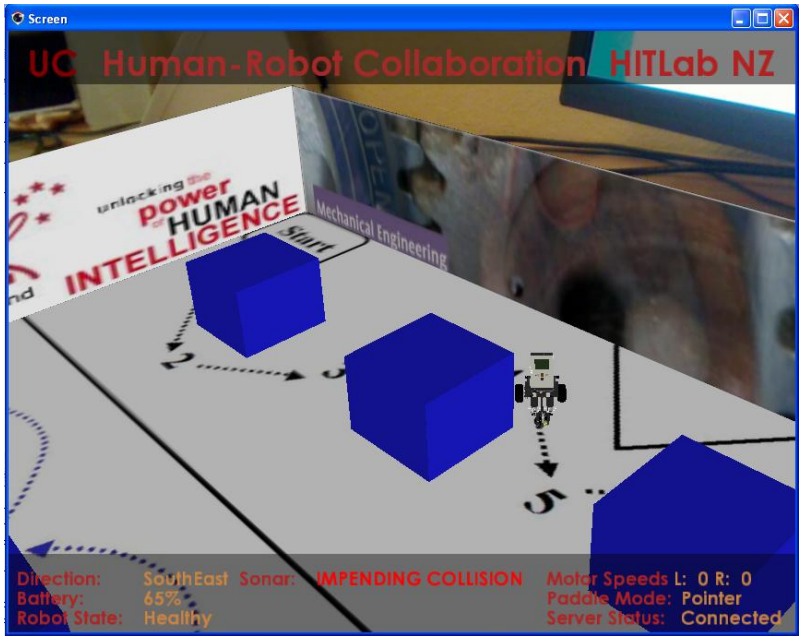


Figure 7.8 Robot determines that something may be in the way and stops. The robot then requests help from human before continuing. The human is able to garner a different perspective of the robot's environment by moving the real world fiducial grid.

Human: *Please continue.*

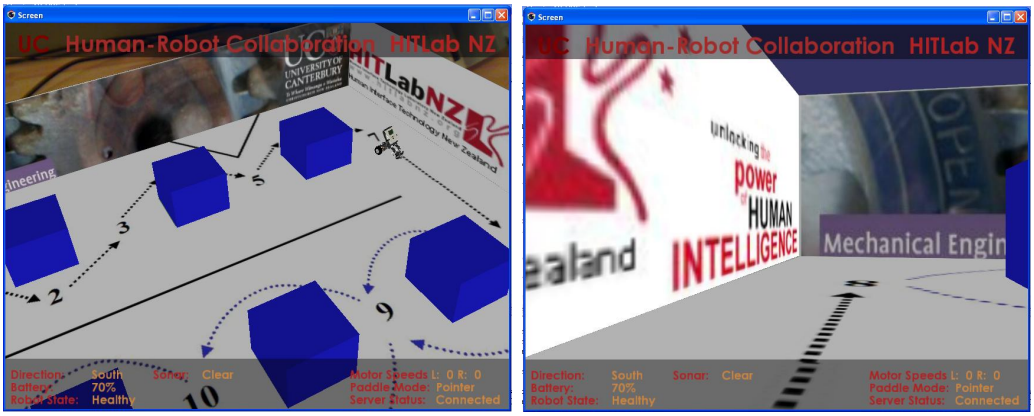
Robot: *Continuing with plan then.*

The robot continues with the plan as it was originally created. However, if the human determines that the robot's path is no longer clear, then the human can insert additional via points to avoid the object, as took place during the creation of the plan. The human can also see what is happening through the eyes of the robot by instructing the AR system to display the world as seen through the robots camera.

Human: *Change to ego view.*

Robot: *Changing to ego view.*

The view presented to the user in the HMD smoothly transitions from an exo-centric (God's eye) view to that seen by the robot through its on-board camera. In this manner, the human is better able to understand what the robot is experiencing from its point of view. The human can easily transition back to the exo-view as well. Figure 7.9 shows both an ego and exo centric view of the robots work space that the user sees in the HMD.



(a) Exo-Centric View (b) Ego-Centric View

Figure 7.9 Exo-centric view of robots workspace (a) and ego-centric view (b).

Human: *Change to exo view.*

Robot: *Changing to exo view.*

The robot then continues on with the plan. When the goal location has been reached the robot alerts the user that the plan has been completed.

Robot: *Arrived at goal location.*

7.4 Summary

This chapter described the integration of a mobile robot into the AR-HRC system. Interaction techniques were discussed and a detailed example of how a user would collaborate with the robot was provided. The next step is to test this experimental instantiation.

Chapter 8

System Evaluation

This chapter provides an evaluation of the AR-HRC system. A user study was conducted where the task involved was to guide a simulated mobile robot through a predefined maze. Three user interfaces were compared for performance and collaboration.

One interface was a typical teleoperation mode with a single ego-centric camera feed from the robot. A second interface was a limited version of the AR-HRC system that allowed the user to see the robot in its work environment through the AR interface, but did not provide any means of pre-planning or review of the robot's intended actions. The third interface was the full AR-HRC system that allowed the user to view the robot in the AR environment and to use spoken dialog and gestures to work with the robot to create and review a plan prior to execution.

The dependent variables measured in the experiments were the time to completion, accuracy in reaching predefined points in the maze, and the number of impending collisions with objects. In addition, the dialog used throughout the experiment was analyzed. Subjective questionnaires were administered after each of the three trials, along with a final questionnaire upon completion of the entire experiment comparing the three interfaces tested.

8.1 Experimental Design

The task for the user study was to work with a simulated robot based on the robot of Chapter 7, and to guide it through a predefined maze. Three

conditions were used:

- ***Immersive Test:*** A typical teleoperation mode with a single egocentric view from the robot's onboard camera.
- ***Speech and Gesture no Planning (SGnoP):*** A limited version of the AR-HRC system that allowed the user to see the robot in its work environment in AR and interact with the robot using speech and gesture, but without pre-planning and review of the robot's intended actions.
- ***Speech and Gesture with Planning, Review and Modification (SGwPRM):*** The full AR-HRC system that allowed the human to view the robot in the AR environment, use spoken dialog and gestures to work with the robot to create a plan and review this plan prior to execution.

The Immersive condition was intended to mimic the traditional teleoperational control of a mobile robot. The SGnoP condition was to introduce a part of the AR-HRC system, namely the ability to reach into the world of the robot through the AR graphics and use spatial dialog to issue commands. The SGwPRM condition also included the AR interaction, but added deeper dialog with the robot and the ability to create and review a plan with the robot prior to its execution. The intent was to see if the SGwPRM condition provided the user a feeling of presence in the robot's world and a feeling of the robot being a collaborative partner rather than a tool. The HUD discussed in Chapter 7 was part of all three interface conditions.

The three conditions are, therefore, distinguished by increasing levels of collaboration or communication channels. Table 8.1 shows the input and output channels of the robot for each condition of the experiment.

Condition	**Input to Robot**	**Output from Robot**
Immersive	Keyboard Input	Ego view of robot
SGnoP	Speech and Paddle Gesture	Exo view of robot work space
SGwPRM	Speech and Paddle Gesture	Exo view of robot work space, Overlay of robot path plan, Verbal responses

Table 8.1 Communication channels for the virtual robot.

8.2 Participants

Ten participants were run through the experiment, seven male and three female. Ages ranged from 28 to 80 and all participants were working professionals. Six participants had Bachelor's degrees and four advanced degrees. Seven of the participants were engineers, while the other three had non-scientific backgrounds. Overall, the users rated themselves as not familiar with robotic systems, speech systems, or Augmented Reality.

8.3 Procedure

The first step of the experiment was to have each participant fill out a demographic questionnaire to evaluate their familiarity with AR, game playing experience, age, gender and educational experience. Since speech recognition was an integral part of the experiment it was necessary to have each participant run through a speech training exercise. This training created a profile for each user so that the system was better able to adapt to the speech patterns of the individual participant.

The objective of each trial was then explained to the participants. They were told that they would be interacting with a virtual mobile robot to get it through a predefined maze. The maze contained a defined path for the robot to follow and various obstacles the robot would need to maneuver around. The maze is shown in Figure 8.1.

The black lines indicate a path that needed to be followed, while the blue lines indicate that the user had the choice of which path to take. The participants were told that the robot must arrive at each of the numbers on the map as a measure of accuracy for the test. Other parameters measured were impending collisions and time to completion.

It was explained to the participants that the robot was located remotely. The effect on the trials of this approach was that when the robot was directly driven, a time delay would be experienced. Thus, a delay in reaction of the simulated robot was not the system failing, but was the result of the time taken for the commands to reach the robot and the update from the robot to

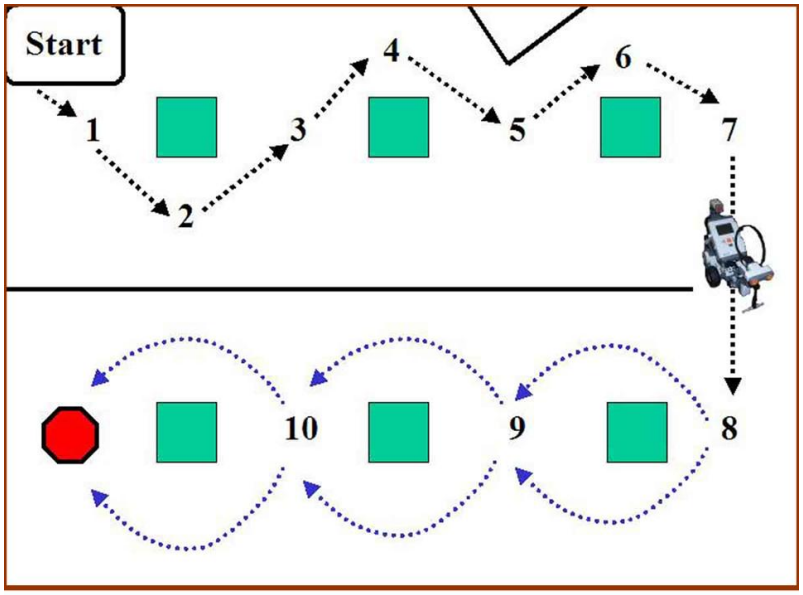


Figure 8.1 The maze used for the user study task. The black lines indicate a defined path to be followed, the blue lines indicate the user has the choice of how to proceed.

arrive back to the user. This detail added a measure of reality to the study.

The experimental setup used was a typical video see through AR configuration. A webcam attached to an eMagin Z800 Head Mounted Display (HMD) (eMagin, 2008) were both connected to a laptop PC running ARToolKit based software. Vision techniques were use to identify unique ARToolKit markers in the user's view and align the 3D virtual images of the robot in its world to these markers. This augmented view was presented to the user in the HMD. Figure 8.2 shows a participant using the AR-HRC system during the experiment.

The same sequence of events took place for each trial. Before the trial was run, the participant practiced using the system to become familiar with the interface for that particular condition. The user also practiced the speech specific to that trial. Once the user felt comfortable with the interface, the trial was run.

When each trial was complete the user was given a subjective questionnaire to determine if they felt that they had a high level of spatial awareness during the trial. The user was also questioned about whether they felt present in the robot's world and their view of the robot as a partner. The participants were also asked to list what they liked and disliked about the interface. This questionnaire was exactly the same for all three trials.

Figure 8.2 A participant using the AR-HRC system. The image on the monitor is what is being displayed to the user in the HMD. The participants did not use the external monitor, it was used to track the progress of the participants.

At the end of the experiment, after the participant had completed all three trials, a subjective questionnaire was given so the user could compare the three conditions. The post trial questionnaires discussed previously referred only to the trial that had just been completed. The subjective questioning was conducted in this manner to let the user express their feeling of each condition individually and then compare the three conditions upon completion of the full experiment. The order of the trials was randomly selected for each user to eliminate the effects of sequencing in the results. All questionnaires for this experiment can be found in Appendix B.

8.3.1 Immersive Condition

The Immersive Test simulated the direct teleoperation of the robot with visual feedback to the user displaying the view that the robot saw through its camera. This view provided the user an ego-centric view of the robot's environment. User interaction included keyed input for robot translation and rotation. The four arrows keys were mapped to forward and backward motion, as well as rotation in the left and right directions. The view the user experienced can be seen in Figure 8.3.

Figure 8.3 The user's view for the Immersive condition. The view shown is that from the robot.

8.3.2 Speech and Gesture no Planning

The SGnoP condition provided the user with a 3D graphic of the robot and maze. The participant was able to use spatial dialog coupled with gestures using a paddle to interact with the graphical world of the robot in the AR environment. Thus, the participant was able to point to a 3D location on the maze and instruct the robot to "go there" or select an object and instruct the robot to "go to the right of that". The robot responded immediately to the verbal commands given, minus the built in time delay for the simulation of a remotely located robot. This time delay was typically on the order of 1 -2 seconds. The view provided to the participant can be seen in Figure 8.4

8.3.3 Speech and Gesture with Planning, Review and Modification

This condition included all the features of the SGnoP condition, but also allowed the participant to use spatial dialog to create a plan with the robot. The user was able to select a goal location then assign waypoints for the robot to follow to arrive at the goal destination. The user could also interactively mod-

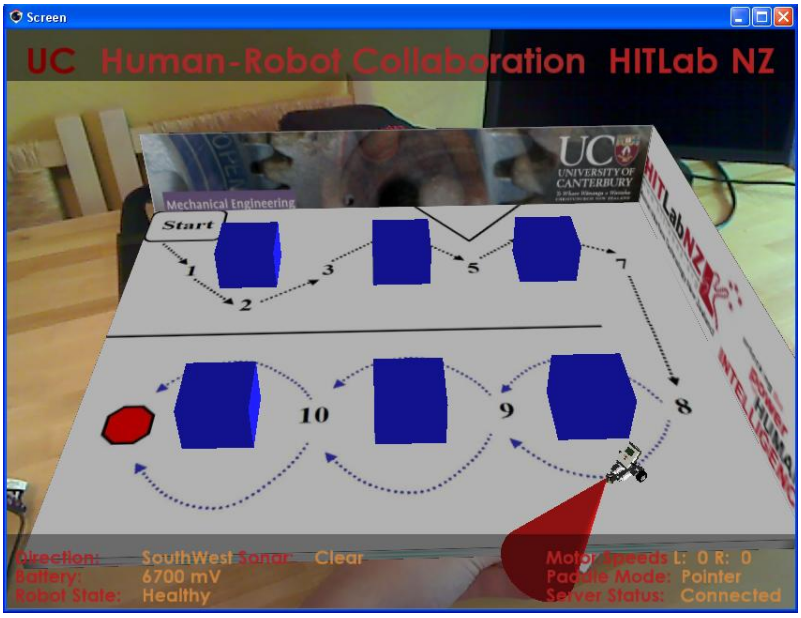


Figure 8.4 The user's view for the Speech and Gesture no Planning condition.

ify the plan by adding or deleting way points. The plan was displayed to the user in the AR environment, thus allowing the participant to determine if the intentions of the robot matched those of the user before any commands were executed by the robot. The robot participated in the dialog by responding to the user verbally for each interaction and verbally alerting the user when the robot came close enough to an object that the robot "thought" it would collide. The user's view for the SGwPRM condition is shown in Figure 8.5.

8.4 Results

The ten participants each performed three tasks, one for each condition. Table 8.2 shows the order in which the ten participants experienced the three interface conditions. Each trial yielded a measure of time to completion, impending collisions and accuracy in reaching each of the ten defined locations on the map. An impending collision was defined as any time the robot came within a predefined threshold of an object. A warning was given to the user that an object too close to the robot, and that a human perspective was needed to determine if the current course of action was clear. The following section reports the results of these measures.

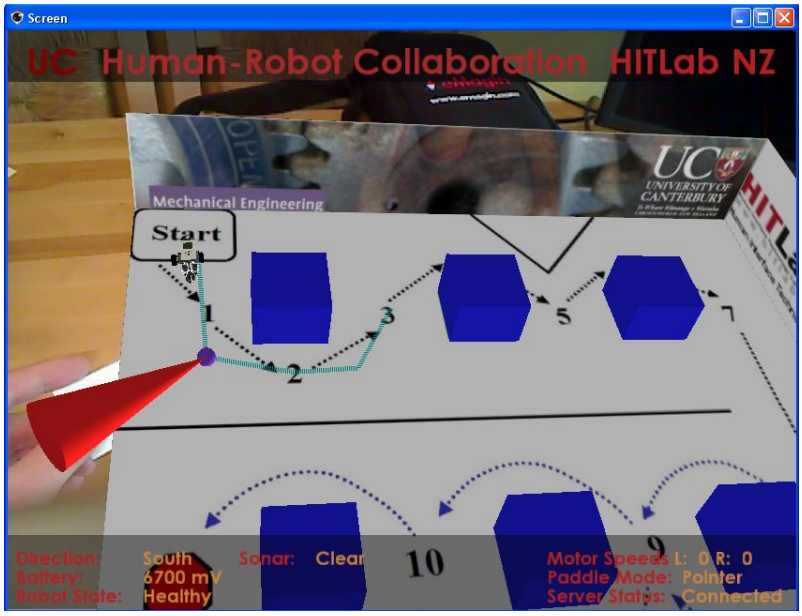


Figure 8.5 The user's view for the Speech and Dialog with Planning, Review and Modification condition. The user creating a plan (blue line) that includes various waypoints through the use of spatial dialog and gesture.

	Trial 1	**Trial 2**	**Trial 3**
Participant 1	Immersive	SGnoP	SGwPRM
Participant 2	SGnoP	SGwPRM	Immersive
Participant 3	Immersive	SGwPRM	SGnoP
Participant 4	SGwPRM	SGnoP	Immersive
Participant 5	SGnoP	Immersive	SGwPRM
Participant 6	SGwPRM	SGnoP	Immersive
Participant 7	Immersive	SGnoP	SGwPRM
Participant 8	SGnoP	SGwPRM	Immersive
Participant 9	SGwPRM	Immersive	SGnoP
Participant 10	Immersive	SGwPRM	SGnoP

Table 8.2 The sequence in which each participant experienced the three interface conditions.

8.4.1 Objective Measures

There was a significant main effect of condition on task completion times with an ANOVA test finding ($F_{2,27} = 9.83$, $p < 0.05$). Bonferroni correction (NIST, 2008) identifies which means are significantly different, and is used in this analysis when the ANOVA test shows a significant main effect of experiment condition. Pairwise comparison with Bonferroni correction ($p < 0.05$) revealed

significant differences between the SGwPRM and the other two conditions while there was no significant difference between SGnoP and the Immersive conditions. The SGwPRM condition was significantly slower than the other two conditions. The results for mean time to completion are shown in Figure 8.6.

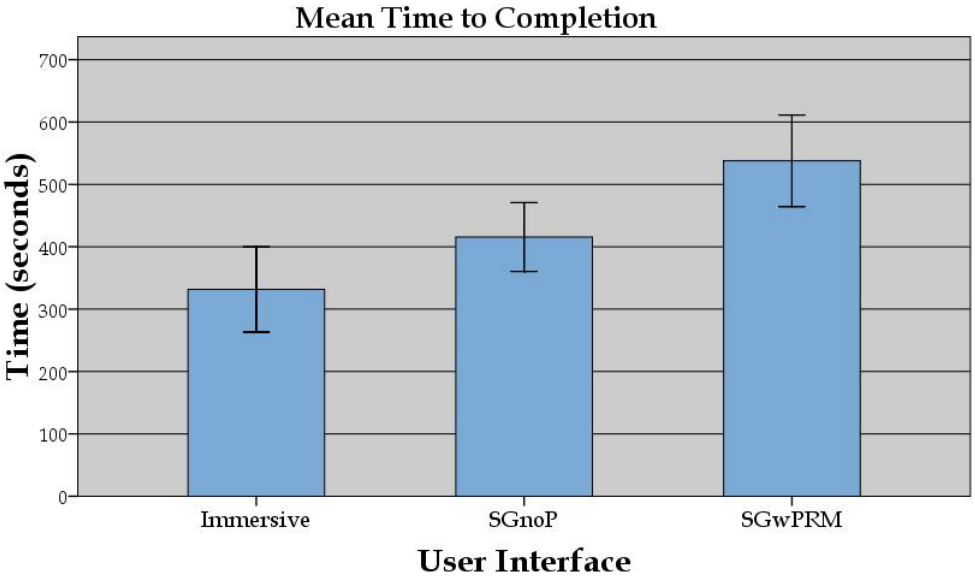


Figure 8.6 Mean time to completion.

The interface condition also had a significant main effect on accuracy with an ANOVA test finding ($F_{2,27} = 8.44$, $p < 0.05$). Accuracy was a count of the number of predefined locations reached during the traversal of the maze. Pairwise comparison with Bonferroni correction ($p < 0.05$) revealed significant differences between the SGwPRM and Immersive conditions, but no significant differences between the SGnoP and the other two conditions. Users in the SGwPRM condition performed the best by arriving at an average of 9.5 out of 10 defined locations (SE = 0.22), and although this result was significantly better than the Immersive condition, there was no significant difference from the SGnoP condition. The results of accuracy measures are shown in Figure 8.7.

There was a significant main effect of condition on the number of close calls with an ANOVA result of ($F_{2,27} = 13.10$, $p < 0.05$). Pairwise comparison using Bonferroni correction ($p < 0.05$) showed significant differences for close calls between the Immersive condition and the other two conditions. There

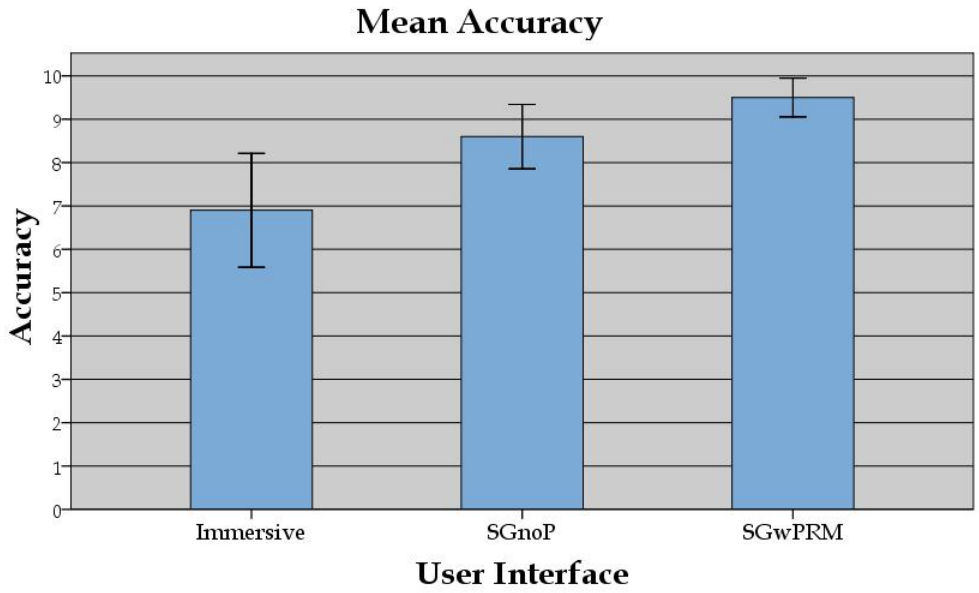


Figure 8.7 Mean accuracy. The graph represents the number of goal locations reached. The maximum was 10.

was no significant difference between SGnoP and SGwPRM. The SGwPRM condition performed best with a mean number of close calls of 3.60 (SE = 1.01), and significantly better than the Immersive condition, although there was no significant difference from the SGnoP condition. The results of close calls are shown in Figure 8.8.

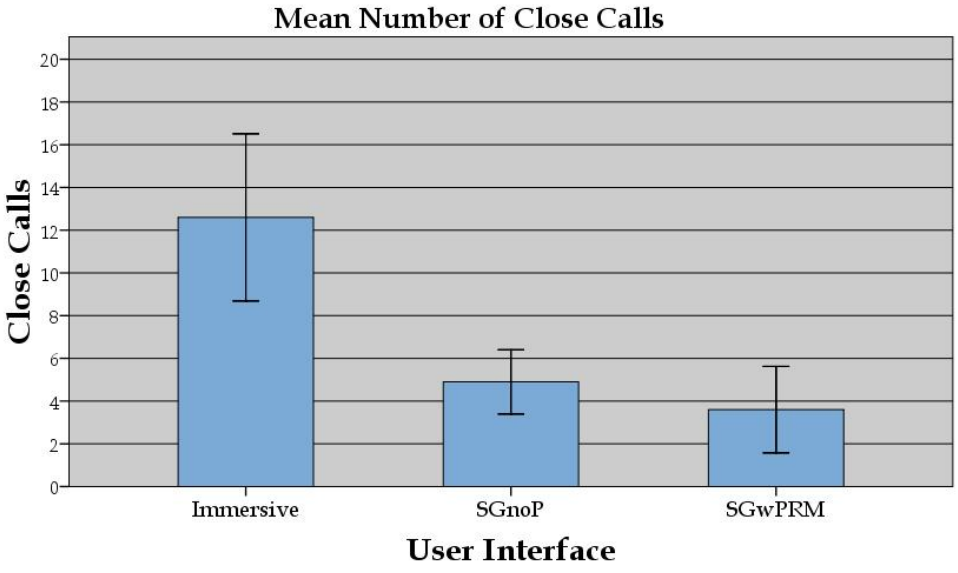


Figure 8.8 Mean number of close calls.

8.4.2 Subjective Measures

The answer for each post trial question was given on a Likert scale of 1- 7 (1 = disagree completely, 7 = agree completely) and analyzed using an ANOVA test. Where necessary, post-hoc analysis was performed using Bonferroni correction ($p < 0.05$). The results of the questionnaires for the individual trials (PT) are presented first and can be seen in Figure 8.9.

- PTQ1: *I knew exactly where the robot was in its world at all times.* There was a significant difference between conditions ($F_{2,27} = 7.43$, $p < 0.05$). Pairwise comparison showed a significant effect between the Immersive condition and the other two conditions, but no significant effect between the SGnoP and SGwPRM conditions. Users felt that they maintained situation awareness best in the SGwPRM and SGnoP conditions.

- PTQ2: *The interface was intuitive to use.* There was no significant difference between the conditions, ($F_{2,27} = 0.03$, $p > 0.05$).

- PTQ3: *The robot was a member of my team as we completed the given task.* There was a significant difference between conditions ($F_{2,27} = 6.07$, $p < 0.05$). Pairwise comparison revealed a significant effect between the Immersive condition and the two others. There was no significant difference between the SGnoP and SGwPRM conditions. The users felt that the robot was a member of their team in the SGwPRM and SGnoP conditions.

- PTQ4: *I felt a sense of being present in the robot's world.* There was no significant difference between the conditions, ($F_{2,27} = 0.37$, $p > 0.05$).

- PTQ5: *I was always aware of how close the robot was to objects in its environment.* There was no significant difference between the three conditions, ($F_{2,27} = 1.84$, $p > 0.05$).

- PTQ6: *I felt like the robot was just a tool and not a collaborative partner.* There was a significant difference between conditions ($F_{2,27} = 5.68$, $p < 0.05$). Pairwise comparison revealed a significant effect between the SGwPRM and Immersive conditions. There was no significant effect between the SGnoP and the other two conditions. Users felt that the robot was more of a collaborative partner in the SGwPRM condition than the Immersive condition.

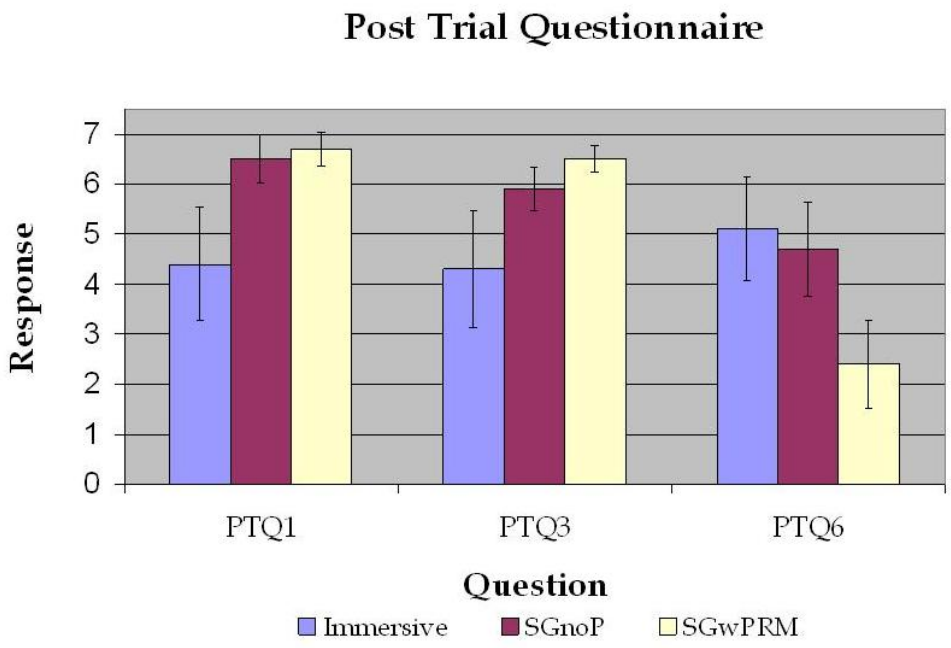


Figure 8.9 Post Trial Questionnaire Results. Likert Scale (1 = Disagree Completely, 7 = Agree Completely)

The results of the post experiment (PE) questionnaire are now presented. As opposed to the questions above which were completed for each condition individually, the users ranked the three conditions in order of preference for the following questions. The results of the post experiment questionnaire can be seen in Figure 8.10.

- PEQ1: *I was aware of collisions as they happened.* There was a significant difference between conditions ($F_{2,27} = 12.47$, $p < 0.05$). Pairwise comparison revealed a significant effect between the SGwPRM and the other two conditions, but no significant effect between the SGnoP and the Immersive conditions. Users felt that they were most aware of collisions while using the SGwPRM condition compared to the other conditions.

- PEQ2: *I had a feeling of working in a collaborative environment.* There was a significant difference between conditions ($F_{2,27} = 17.90$, $p < 0.05$). Pairwise comparison revealed a significant main effect between SGwPRM and the other two conditions, but no significant effect between the Immersive and SGnoP conditions. The SGwPRM condition was selected as providing the users with the greatest feeling of working in a collaborative environment.

- PEQ3: *I felt the robot was a partner.* There was a significant difference between conditions ($F_{2,27} = 17.90$, $p < 0.05$). Pairwise comparison revealed a significant main effect between SGwPRM and the other two conditions, but no significant effect between the Immersive and SGnoP conditions. The SGwPRM condition provided the users with a feeling that the robot was a partner.

- PEQ4: *The interface was intuitive to use.* There was no significant difference due to condition, ($F_{2,27} = 2.28$, $p > 0.05$).

- PEQ5: *I was aware of the robot's surroundings.* There was a significant difference between conditions ($F_{2,27} = 8.39$, $p < 0.05$). Pairwise comparison showed a significant effect between the SGwPRM and Immersive conditions, but no significant effect between the SGnoP and the other two conditions. Users felt that the SGwPRM condition enabled them to be more aware of the robot's surroundings compared to the Immersive condition, but not the SGnoP condition.

- PEQ6: *I had to always pay attention to the robot's actions.* There was a significant difference between conditions ($F_{2,27} = 8.77$, $p < 0.05$). Pairwise comparison showed a significant effect between the Immersive condition and the two others, but no significant effect between the SGnoP and SGwPRM conditions. User felt that they needed to pay attention to the robot's action more in the Immersive condition than the other two conditions.

- PEQ7: *I felt the robot was a tool.* There was no significant difference between the three conditions, ($F_{2,27} = 0.42$, $p > 0.05$).

- PEQ8: *I felt I was present in the robot's environment.* No significant difference was found between the three conditions, ($F_{2,27} = 0.36$, $p > 0.05$).

- PEQ9: *I knew when the robot was about to collide with an object.* There was a significant difference between conditions ($F_{2,27} = 9.62$, $p < 0.05$). Pairwise comparison revealed a significant effect between the SGwPRM and the other two conditions, but no significant difference between the Immersive and SGnoP conditions. Participants felt that the SGwPRM condition was best for maintaining awareness of potential collisions compared to the other two conditions.

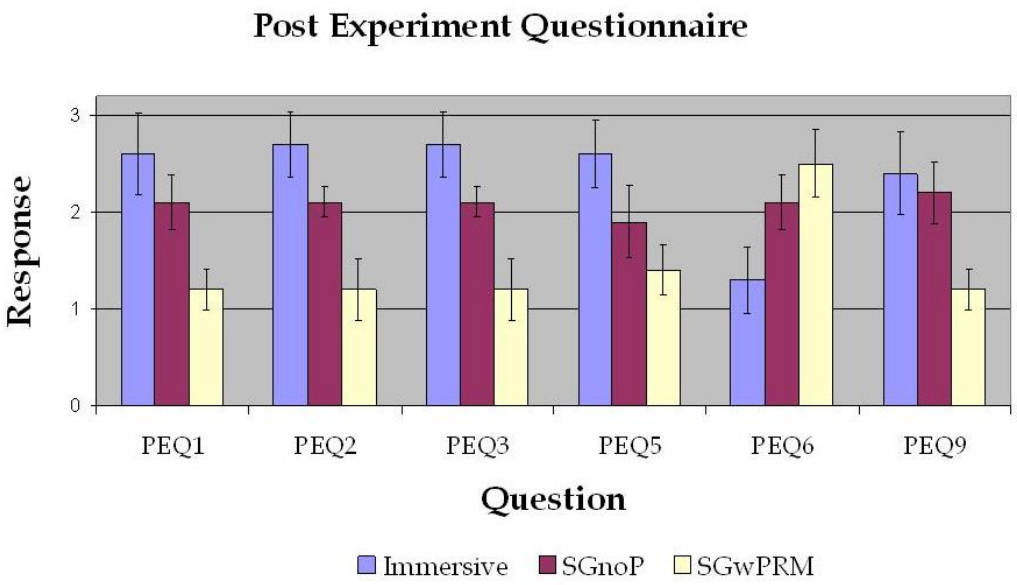


Figure 8.10 Post Experiment Questionnaire Results. 1 = Most Preferred, 3 = Least Preferred

8.4.3 Participant Comments

Users were asked to comment on each of the three interface conditions. Participants commented that for the Immersive condition they liked how the interface was simple and straight forward to use and thus there was no learning curve involved. For the Immersive condition, users commented they did not like the limited view from the robot and the lack of feedback. They also felt it was much harder to succeed in the Immersive condition for the reasons just given.

Similarly for the SGnoP condition, users commented that they liked the dialog with the robot, the ability to pick up and manipulate the virtual representation of the robot's workspace and the ability to view the robot's progression from an exo-centric viewpoint. Participants commented that it was challenging to use the pointer in the AR environment and that they had difficulty remembering the verbal commands.

Finally, for the SGwPRM condition, users commented that they liked the interactive plan creation and modification. Additionally, users commented that they liked the ability to change the perspective by moving the fiducial marker grid and the interactive dialog with the system. However, two users commented that the use of the HMD made them dizzy. Users also commented

that it took time to become accustom to the interaction in the AR environment and that they had difficulty remembering the verbal commands.

8.5 Discussion

The Immersive condition was significantly faster than the SGwPRM condition. This result could be in part due to the lower learning curve of the Immersive condition. This hypothesis is supported by comments users provided in the post experiment questionnaire. Five users commented that the Immersive condition was simple and straight forward to use or that there was no learning curve. In contrast, the SGwPRM condition was a bit more difficult for the participants to become acquainted with.

This higher learning curve has two main causes. First, the user had to become familiar with the dialog that the system understood in a relatively short period of time. Second, at the same time, the users also had to become familiar with selecting locations and objects in the AR environment.

In the Immersive condition, the participants did complete the task faster than the SGwPRM condition. However, the measure of accuracy showed that the users performed worse in the Immersive condition compared to the SGwPRM condition. The participants performed best in terms of accuracy in the SGwPRM condition as opposed to the Immersive condition. So although the SGwPRM condition took on average the longest time to complete the task, it resulted in the more accurate performance compared to the Immersive condition.

It is not surprising to see that the SGwPRM has a longer completion time. This result is inherent in the design of the interface as it takes time for the robot to display its plan in AR, for the user to agree with or modify the plan, and then have the robot execute the plan. Thus, greater planning leads to better outcome as might be expected with a good collaboration environment.

There was a significant effect on the number of close calls. The condition that performed the worst in this measure was the Immersive condition. This result combined with the results from questions PTQ1, PEQ1, PEQ5 and PEQ9 indicate that the SGwPRM condition provided the users with the

highest level of spatial awareness.

An analysis of the dialog used revealed that deictic phrases, such as "go here", were used 87% of the time for the SGnoP condition and 93% of the time for SGwPRM. The remaining times deeper spatial dialog was used, such as "to the left of this" while selecting an object in the AR environment. This result of mainly using the deictic gestures could be due to the learning curve mentioned previously.

In particular, to use the deeper spatial dialog the participants had to remember longer phrases and coordinate issuing these phrases with the selection of objects in AR. Although this coordination is not difficult to master with practice, the participants tended to use a method that they could immediately master. The use of the deeper spatial dialog thus tended to happen later in the experiment, once the participants had become familiar with interacting with the system.

Another subjective measure was the feeling of working in a collaborative environment. The responses from questions PTQ6, PEQ2 and PEQ6 show that the users felt that they were working in a collaborative environment when completing the task using the SGwPRM condition. Question PEQ3 responses show that participants felt the robot was a partner when working with in the SGwPRM condition. These results show that participants felt they were working in a collaborative team environment in the SGwPRM condition.

The last subjective question posed to the users was to select the most effective condition. Nine of the ten participants selected the SGwPRM as the most effective. The remaining user selected the SGnoP condition. Reasons provided for the selection of SGwPRM included effective path creation, verbal feedback from the robot and the ability to change the plan mid-stream. Conversely, reasons given for not choosing the other conditions included the lack of planning caused crashes, the Immersive condition lacked situation awareness and limited feedback from the robot. These results show that being able to exchange dialog with the robot and seeing the robot's intentions does indeed create a collaborative environment.

8.6 Summary

This chapter presented an experiment conducted to evaluate the AR-HRC system. The experiment involved using three interfaces for working with a remotely located mobile robot. One interface was direct teleoperation where the user received visual cues from a camera mounted on the robot and drove the robot through direct teleoperation. A second interface provided the user with an exo-centric view of the robot in its work environment and enabled the human to use speech and gesture to communicate to the robot where it was to go.

The third interface provided the user with the same exo-centric view of the robot and allowed for spatial dialog and gesture interaction. However, this interface also enabled the human to collaborate with the robot to create, modify and review a plan before the robot executed it. This interface comprises the Augmented Reality Human-Robot Collaboration System at the centre of this thesis.

Subjective questioning showed that users felt they were working in a collaborative environment when using the AR-HRC interface. In this interface, users also felt that they maintained better situation awareness, which is supported by the objective measurements of accuracy and close calls. Users also felt that the robot was more of a partner in the AR-HRC interface.

The users overwhelmingly selected the AR-HRC interface as the most effective of the three interfaces tested. The results of this study show that by providing the human with a shared view of the robots workspace and enabling the human to use natural speech and gesture, with robotic verbal feedback, effective communication can take place between the robot and human. Common ground is easily reached by visually displaying the robots intentions in this shared workspace. Therefore, an environment has been created that allows for effective communication, and thus, collaboration.

Chapter 9

Conclusions

This thesis leads the reader through the development of the AR-HRC system and approach to human-robot collaboration from concept and background through the design of the necessary set of interfaces required. It thus began by introducing the need for human-robot collaborative teams in terms of current and emerging application spaces requiring collaboration to achieve or significantly improve outcomes. In particular, the area of space exploration will require human-robot interaction at levels well beyond current state of the art or understanding. Similar terrestrial applications are outlined that will be significantly enhanced, as well. However, it was also shown that little attention has been paid to research in this field. All of these issues provided the impetus for the creation of the Augmented Reality Human-Robot Collaboration (AR-HRC) system described here.

A discussion of the related work in HRI has shown that an effective system should transfer the interaction mechanisms natural for humans to the precision required for machine information. Previous work in HRI has also shown that the autonomy level of an HRI system should be variable so that it can match the needs of a given situation. In this manner, the system is able to capitalize on the problem solving skills of a human, while also effectively balancing that with the speed and dexterity of a robot.

Prior work in HRI also highlighted the importance of situation awareness. The lack of situation awareness has been shown to decrease performance and, in certain cases, can lead to catastrophic failures. Use of natural speech has also been shown to be effective in HRI. However, speech alone is not enough to complete the grounding process in the exchange between human and robot,

leading to a reduced ability to communicate as a result. Therefore, a multimodal interface is shown to provide a more effective approach. By combining speech with gesture, a more natural interface and the requisite grounding is achieved. The multimodal medium used for the AR-HRC system presented here is Augmented Reality, which affords both speech and gestural communication channels.

Therefore, the literature review includes an introduction to AR and the state of the art of AR in the context of using it in a multimodal human-robot interaction system. AR has been shown to provide a shared work space that is conducive to collaboration and at the same time increases situation awareness, enhancing its potential in this situation. AR also supports a tangible user interface, essentially allowing a person to use a real world object to affect change on the 3D graphics of the AR environment, providing an enhanced graphical or visual communication channel. AR was also shown to increase performance in robotic control directly. In particular, the use of AR improved situation awareness by providing the human with an exo-centric view of the robots workspace. Therefore, AR provides rich spatial cues in the shared environment and enables the use of natural spatial dialog. By taking explicit advantage of the benefits that AR offers, a robust human-robot collaboration system can be created.

As a first step towards the development of the AR-HRC system, a multimodal interface for AR was created. This interface fused spatial dialog and gesture interaction to affect change in an AR environment. The results of a user study for this system showed that the multimodal interface improved performance in the AR environment. These positive results drove the design of the AR-HRC system to include multimodal AR interaction through the use of spatial dialog and gestures.

The architectural design of the AR-HRC system was then presented. The various components of the system were described in detail. The intercommunication of these modules was also discussed. The system design is seen to fuse speech and gesture inputs with the AR overlays of the robots plans and internal state. As a result, the system is able to provide a communication environment this is equally and highly effective for both parties in the human-robot collaboration.

The thesis then discussed a Wizard of OZ study conducted that helped to define the the type of speech and gesture interaction to incorporate into the AR-HRC system. Given the opportunity, participants used natural speech and gestures to work with a robotic team member. Initially, with no instructions given on what type of speech and gestures to use, the participants communicated with the robot in a manner they thought the robot would understand. This manner was short, mechanized terminology. However, once the participants learned they could communicate in a natural fashion they did so and commented on the natural and intuitive nature of the interface.

It was observed that when participants used more descriptive communication behaviour, the result was fluid robot motion and reduced completion times. Participants also commented on the usefulness of having the robot verbally respond to enable them to maintain awareness of what the robot was doing and what it was "thinking". Therefore, a multimodal approach to human-robot communication results in the most effective communication taking place, thus enhancing the collaborative interaction.

The integration of a mobile robot into the AR-HRC system was then presented. The environment the robot was to work in was described, as well as a task for the robot to complete. The ability to create, review and modify robot plans was described highlighting the collaborative nature of the AR-HRC system.

A performance experiment comparing three user interfaces was then discussed. The three interfaces used were:

- A typical teloperation interface
- A version of the AR-HRC that did not include planning or review
- The full version of the AR-HRC that did include path planning, review and modification

Each of these interfaces was described in detail. The task to be completed, the variables measured and the subjective questionnaires participants filled out were also discussed. Results showed that participants felt the robot was more of a tool in the teleoperation interface. Participants thought of the robot as

more of a collaborative partner when using the full version of the AR-HRC interface.

The ability of the system to immerse the user in the remote environment of the robotic system resulted in the users perceiving the interaction as a collaborative one. This feeling of telepresence resulted in the participants of the evaluation study perceiving the robot as a collaborative partner and not as a tool, as robots are typically perceived. The overall impact is that by providing the feeling of telepresence, the AR-HRC system allows the human to overcome the initial tendency to treat the robot as a mechanical device and the interaction to evolve into a more natural team-centric collaborative one.

While these results might be as expected, they clearly highlight the change in perception of the human partner in the robots capability that arises with increasingly effective two-way communication through an environment explicitly designed to maximize that collaborative discussion. Hence, it is clear that human-robot interaction, while a nascent field, can offer significantly improved task performance for both robot and operator, even in the simple proof of concept studies presented here. Thus, the main conclusion of this thesis is that human-robot collaboration represents an immediate and significant frontier to be crossed on the way to developing next generation robotic applications and that AR technology can be of significant benefit in this work.

The development and evaluation of the AR-HRC system took a multidisciplinary approach. An engineering approach was taken in the design of the system. In addition, a subjective approach was also integrated in the development of the system from the results of the user studies. The Wizard of OZ study in particular was designed to include the ideas and interactions of real users with the system before the system was fully developed. This approach has resulted in a system that truly reflects the desired interaction techniques of users of the system.

In summary, this thesis has shown that the AR-HRC system concept does enable natural and effective communication to take place. The use of AR affords the integration of a multimodal interface combining speech and gesture interaction, as well as providing the means for enhanced situation awareness. The AR-HRC system gives the user the feeling of working in a collaborative human-robot team rather than the feeling of the robot being a tool, as a typical

teleoperation interface provides. Therefore, the development of the AR-HRC system brings closer the day when humans and robots can truly interact in a collaborative manner, as well as highlighting the main requirement all such systems must meet to ensure such quality in collaboration.

Chapter 10

Future Work

The AR-HRC system presented in this thesis can be viewed as a first step into an emerging research area in Human-Robot Interaction, namely that of multimodal interactive collaboration with robotic systems. With that in mind, there exists opportunities to expand on this research. These opportunities are presented first by modules of the AR-HRC system, then the system as a whole and finally some potential areas for integration and evaluation studies.

10.1 AR-HRC Modules

Speech recognition and text-to-speech obviously play a major role in the AR-HRC system and are themselves an active field of research. As this field matures further, false detection rates will be reduced and, consequently, recognition rates will increase. As false detection rates are reduced it will be possible to create dialog that more closely replicates how humans speak. One way that accuracy of speech in put can be improved is by defining more complex phrases for a situation than may be necessary. For example, instead of having a command of just "stop", the AR-HRC system uses "robot stop". The word "robot" was added to the goal phrase to prevent the system from falsely recognizing the single syllable word "stop" from other utterances of the user or background noise. Important research could also be conducted on the optimum speech grammar for HRI.

The AR-HRC system uses Microsoft Speech for text-to-speech feedback. The options for voice selection are limited and sound very robotic. The implementation of a commercial speech synthesis system might offer more options

for less robotic sounding voices. The intent of the research presented in this thesis was not to explore speech recognition or text-to-speech, but to incorporate this technology into the AR-HRC. Therefore, an avenue for future research would be an improved speech recognition and text-to-speech package, particularly one with a greater range of flexibility to enhance communication channels between human and robot.

Augmented Reality is another active field of research. There are numerous avenues being pursued to enhance AR technology, a few are listed here:

- Outdoor tracking
- Mobile AR applications
- Natural feature tracking / marker-less tracking
- Reduction of noise in tracker output
- World model creation

Improvements in AR outdoor tracking, mobile AR applications and natural feature tracking would provide the ability to take the AR-HRC system out of the laboratory and into the outside world. On-the-fly world model creation would enable the system to be used in new unmapped environments by eliminating the need to create the virtual world prior to using the system. These enhancements combined would enable the system to be effectively used in exploration and surveying tasks, or virtually any situation since no pre-mapped environment would be needed and the human would be free to operate outdoors.

10.2 The AR-HRC System

The AR-HRC system could also be enhanced through further research. In particular, a proof of concept application with a mobile robot was described in this thesis, numerous other robotic applications could benefit from the HRI techniques afforded by the AR-HRC system. For example, Lunar or Martian rovers are possible applications for the AR-HRC. Unmanned Aerial Vehicles

(UAVs), Unmanned Underwater Vehicles (UUVs) and terrestrial rovers, to name just few, could also benefit from the HRI techniques presented in this thesis. In addition, with each new application the dialog will need to be catered to that specific domain and a variety of evaluation studies will need to be conducted to determine how best to implement the system to the given application.

Gesture interaction is yet another area of active research. A variety of gesture interaction methods could be explored for use in the AR-HRC system. Data gloves, visual hand tracking, and even the use of the Nintendo WiiTM remotes (Nintendo, 2008) could be explored as gesture input devices. Computer vision based natural hand input is a particularly promising area of current research that could be extended for HRI.

Improvements or variations to the display device could be explored as well. The implementation presented in this thesis used a head mounted display (HMD). Other possibilities include large LCD screens, white boards, or even the use of a Cave Automatic Virtual Environment (CAVE) and fully immersive graphics environments. Research could also be conducted on the impact of display on human-robot communication.

10.3 Integration and Evaluation Studies

The AR-HRC system could also be expanded to accommodate multiple humans and multiple robots. Possible scenarios could include co-located humans or humans located remotely from each other. These groups could be interacting with a single robot or several robots that do not necessarily have to be located in the same work space. The use of the AR-HRC system has the possibility of taking the complex scenario of a single person collaborating with multiple robots and reducing it to a collaborative interaction that puts less cognitive load on the human. This high level interaction of the user is achieved by providing the human a view through the AR overlay of the workspace of the team of robots and by the use of adjustable autonomy, letting each robotic member of the team operate autonomously and interact with the human collaborator when warranted. The human can monitor the progress of the robotic team members and intervene when the human deems it necessary, thus providing a

collaborative environment that includes multiple robotic team members. This scenario is similar to many video games in use today where a single user is interacting with multiple virtual agents and interacts with these agents when the situation requires it.

Appendix A

Wizard of OZ Study Questionnaires

User Study Demographic Information

Gender: M / F

Age Group:

< 25 26–30 31-35 36-40 41-45 46-50 51-55 >55

Education Level:

Profession:

What is your familiarity with robotic systems?

(1 very familiar ←→ 7 not familiar at all)
1 2 3 4 5 6 7

Do you normally tend to use gestures as you speak?

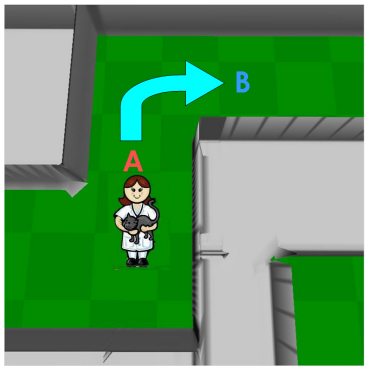
(1 lots of gestures ←→ 7 no gestures)
1 2 3 4 5 6 7

What is your familiarity with speech systems?

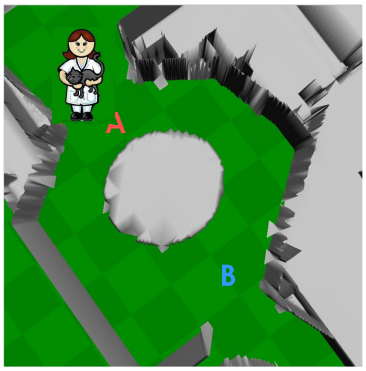
(1 very familiar ←→ 7 not familiar at all)
1 2 3 4 5 6 7

Please describe how you would collaborate with the human to go from A to B. She must stay on the green areas.

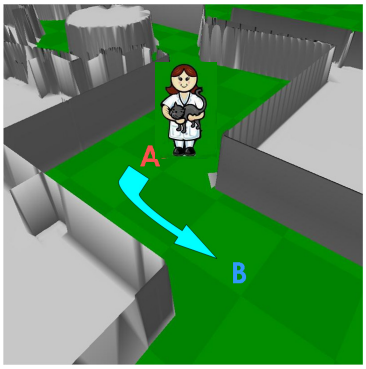
Using SPEECH only.



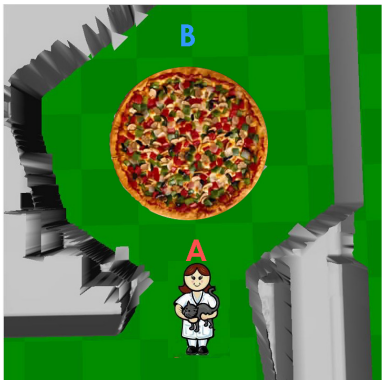
Using SPEECH only.



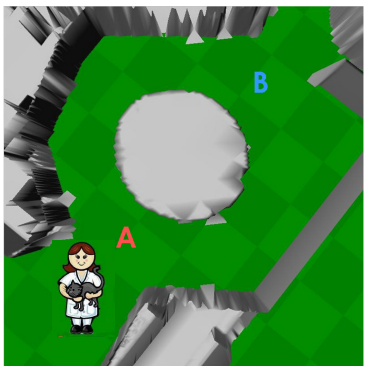
Using SPEECH only.



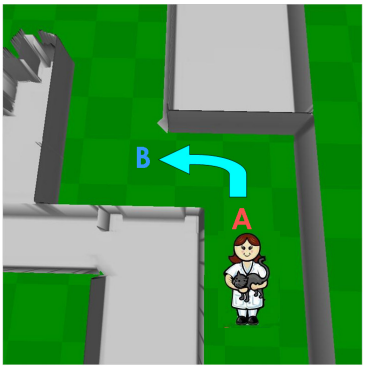
Using SPEECH only.



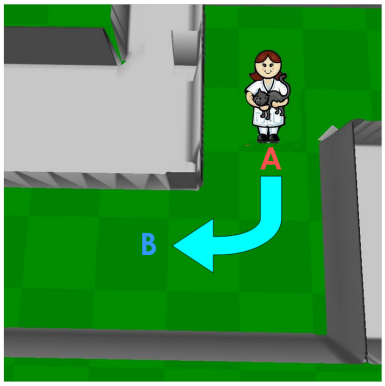
Please describe how you would collaborate with the human to go from A to B. She must stay on the green areas.



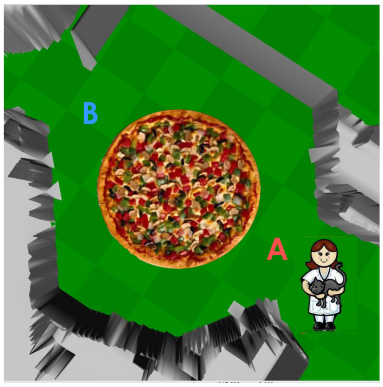
Using GESTURES only.



Using GESTURES only.

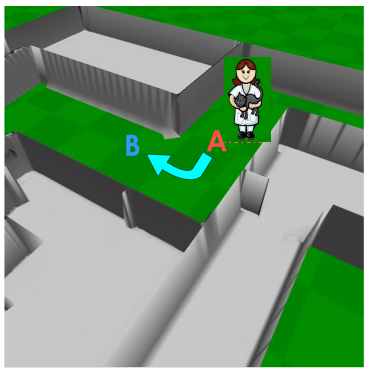


Using GESTURES only.

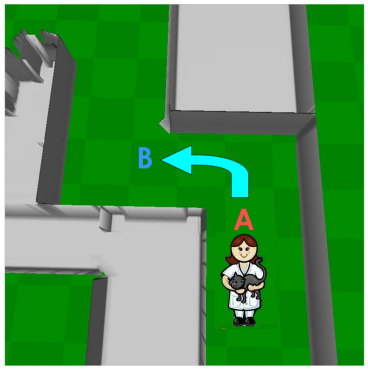


Using GESTURES only.

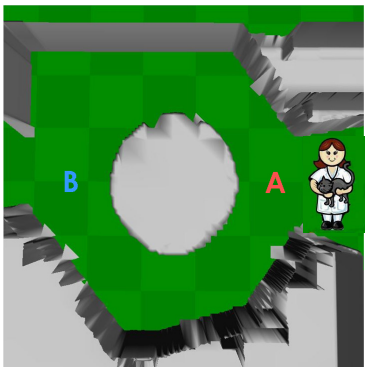
Please describe how you would collaborate with the human to go from A to B. She must stay on the green areas.



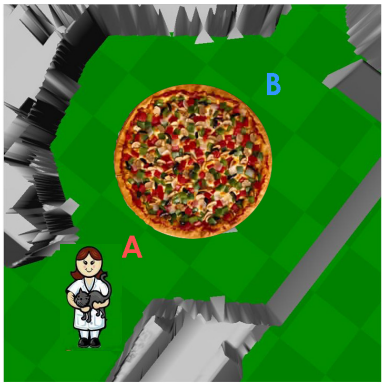
Using combination of speech and gestures.



Using combination of speech and gestures.



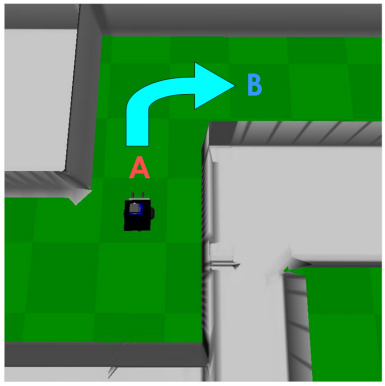
Using combination of speech and gestures.



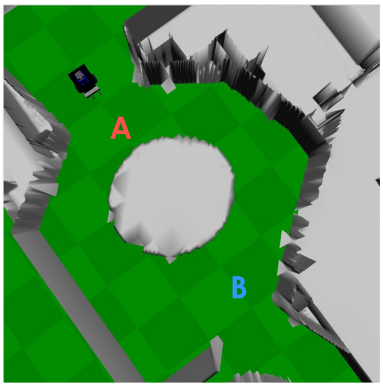
Using combination of speech and gestures.

Please describe how you would collaborate with the robot to go from A to B. The robot must stay on the green areas.

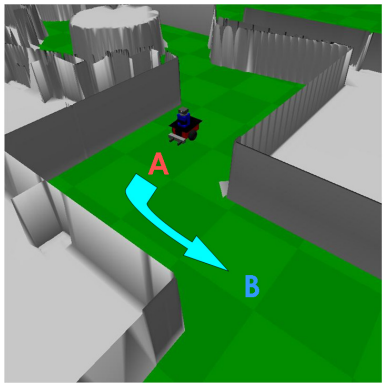
Using SPEECH only.



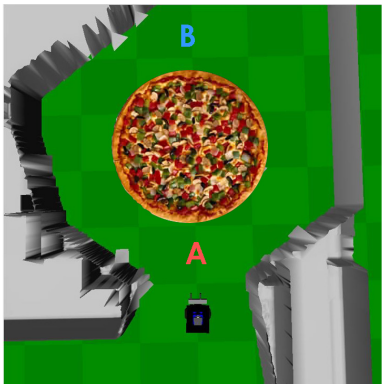
Using SPEECH only.



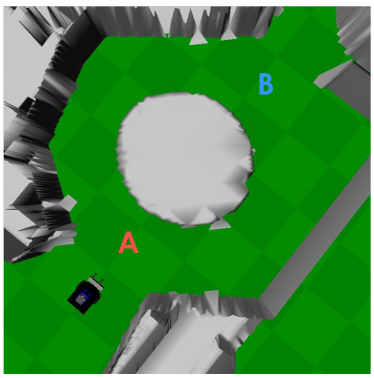
Using SPEECH only.



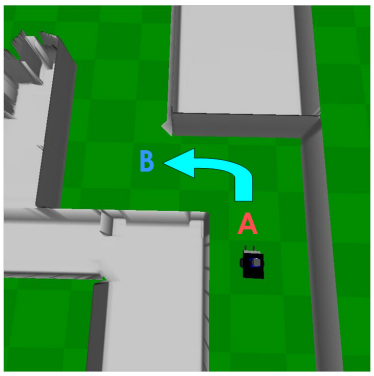
Using SPEECH only.



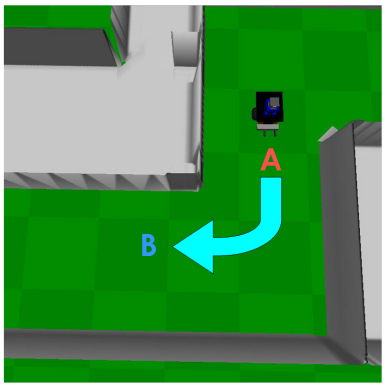
Please describe how you would collaborate with the robot to go from A to B. The robot must stay on the green areas.



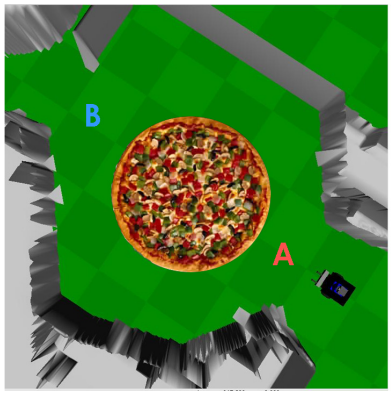
Using GESTURES only.



Using GESTURES only.

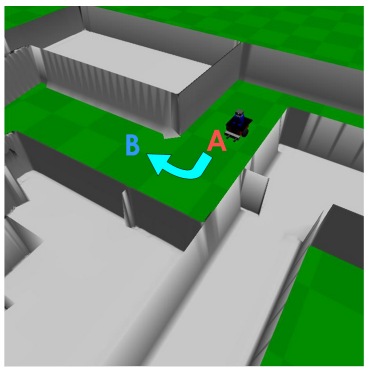


Using GESTURES only.

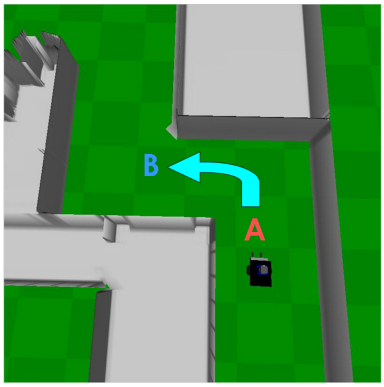


Using GESTURES only.

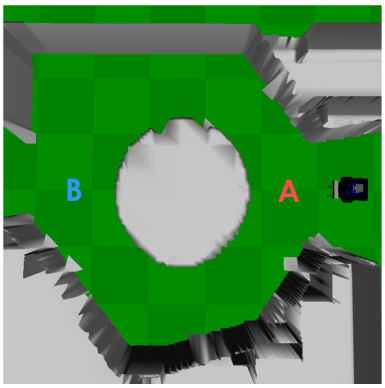
Please describe how you would collaborate with the robot to go from A to B. The robot must stay on the green areas.



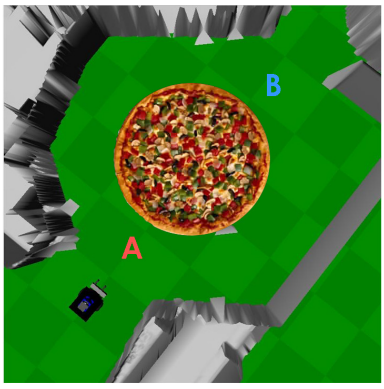
Using combination of speech and gestures.



Using combination of speech and gestures.



Using combination of speech and gestures.



Using combination of speech and gestures.

User Study Post Experiment Questionnaire

(Circle appropriate answer)

Do you feel the system reacted the way you thought it would before you began the experiment?

(1 very much so ←→ 7 not at all)

1 2 3 4 5 6 7

How well did you feel the system understood your verbal spatial references?

(1 very much so ←→ 7 not at all)

1 2 3 4 5 6 7

How well did you feel the system understood the gestures you used?

(1 very much so ←→ 7 not at all)

1 2 3 4 5 6 7

How well did you feel the system reacted the way you wanted it to?

(1 very much so ←→ 7 not at all)

1 2 3 4 5 6 7

Do you feel the use of gestures helped you communicate spatially with the system?

(1 very much so ←→ 7 not at all)

1 2 3 4 5 6 7

Did you have confidence speaking to the system?

(1 very much so ←→ 7 not at all)

1 2 3 4 5 6 7

Did you have confidence gesturing to the system?

(1 very much so ←→ 7 not at all)

1 2 3 4 5 6 7

Do you feel the speech mode was best?

(1 very much so ←→ 7 not at all)

1 2 3 4 5 6 7

Do you feel the combined speech and gesture mode was best?

(1 very much so ←→ 7 not at all)

1 2 3 4 5 6 7

Do you feel the gesture mode was best?

(1 very much so ←→ 7 not at all)

1 2 3 4 5 6 7

Appendix B

Interface Evaluation Questionnaires

Pre-Experiment Questionnaire

Participant:

Gender: M / F

Age Group:

18-25 25-30 30-35 35+

Education Level:

Profession:

How familiar are you with robotic systems?

(1 not familiar at all ← → 7 very familiar)

1 2 3 4 5 6 7

How familiar are you with speech systems?

(1 not familiar at all ← → 7 very familiar)

1 2 3 4 5 6 7

How often do you play videos games?

(1 never ← → 7 all the time)

1 2 3 4 5 6 7

How familiar are you with Augmented Reality?

(1 not familiar at all ← → 7 very familiar)

1 2 3 4 5 6 7

Post-Trial Questionnaire

Subject: **Test:**

(1 Disagree completely ← → 7 Agree completely)

I knew exactly where the robot was in its world at all times.

1 2 3 4 5 6 7

The interface was intuitive to use.

1 2 3 4 5 6 7

The robot was a member of my team as we completed the given task.

1 2 3 4 5 6 7

I felt a sense of being present in the robot's world.

1 2 3 4 5 6 7

I was always aware of how close the robot was to objects in its environment.

1 2 3 4 5 6 7

I felt the robot was just a tool and not a collaborative partner.

1 2 3 4 5 6 7

What I liked about this interface:

What I did not like about this interface:

Post-Experiment Questionnaire

Subject:

Please rank the interfaces for the following attributes: 1 = most preferred, 3 = least preferred	Immersive Test	Speech & Dialog No Planning	Speech & Dialog with Planning
I was aware of collisions as they happened.			
I had a feeling of working in a collaborative environment.			
I felt the robot was a partner.			
The interface was intuitive to use.			
I was aware of the robot's surroundings.			
I had to always pay attention to the robot's actions.			
I felt the robot was a tool.			
I felt I was present in the robot's environment.			
I knew when the robot was about to collide with an object.			

Overall which interface do you feel is the most effective? (Please circle one)

Immersive Interface Speech & Dialog No Planning Speech & Dialog with Planning

Please try to list three reasons why you selected the interface above:

Please try to list three reasons you did not select the other two interfaces:

References

Argyle, M. (1967). The psychology of interpersonal behaviour. London: Penguin Books.

ARToolKit (2008). *www.hitl.washington.edu/artoolkit/, accessed January 2008*.

Azuma, R., Y. Baillot, R. Behringer, S. Feiner, S. Julier, and B. MacIntyre (2001). Recent advances in augmented reality. *IEEE Computer Graphics and Applications 21*(6), 34–47.

Azuma, R. T. (1997). A survey of augmented reality. *Presence: Teleoperators and Virtual Environments 6*(4), 355–385.

Bechar, A. and Y. Edan (2003). Human-robot collaboration for improved target recognition of agricultural robots. *Industrial Robot 30*(5), 432–436.

Bekey, G., R. Ambrose, V. Kumar, D. Lavery, A. Sanderson, B. Wilcox, J. Yuh, and Y. Zheng (2008). In *Robotics: State of the Art and Future Challenges*. Imperial College Press.

Billinghurst, M., J. Bowskill, N. Dyer, and J. Morphett (1998). Spatial information displays on a wearable computer. *IEEE Computer Graphics and Applications 18*(6), 24–31.

Billinghurst, M., R. Grasset, and J. Looser (2005). Designing augmented reality interfaces. *Computer Graphics SIGGRAPH Quarterly 39(1)*, 17–22 Feb.

Billinghurst, M., H. Kato, and I. Poupyrev (2001). The magicbook: A transitional ar interface. *Computers and Graphics (Pergamon) 25*(5), 745–753.

Billinghurst, M., I. Poupyrev, H. Kato, and R. May (2000). Mixing realities in shared space: An augmented reality interface for collaborative computing.

In *2000 IEEE International Conference on Multimedia and Expo (ICME 2000), Jul 30-Aug 2*, New York, NY.

Bolt, R. A. (1980). Put-that-there: Voice and gesture at the graphics interface. *Proceedings of the International Conference on Computer Graphics and Interactive Techniques 14*, 262–270.

Breazeal, C. (2004). Social interactions in HRI: The robot view. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews Human-Robot Interactions 34*(2), 181–186.

Breazeal, C., A. Brooks, J. Gray, G. Hoffmann, C. Kidd, H. Lee, J. Lieberman, A. Lockerd, and D. Mulanda (2003). Humanoid robots as cooperative partners for people. *MIT Media Lab, Robotic Life Group, International Journal of Humanoid Robots December 15*.

Breazeal, C., A. Edsinger, P. Fitzpatrick, and B. Scassellati (2001). Active vision for sociable robots. *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans 31*(5), 443–453.

Burke, J., R. R. Murphy, M. D. Coovert, and D. L. Riddle (2004). Moonlight in Miami: Field study of human-robot interaction in the context of an urban search and rescue disaster response training exercise. *Human Computer Interaction 19*, 85 – 116.

Carbini, S., L. Delphin-Poulat, L. Perron, and J. E. Viallet (2006). From a Wizard of Oz experiment to a real time speech and gesture multimodal interface. *Signal Processing Journal: Special Issue on MultiModal Human-Computer Interfaces 86*(12), 3559 – 3557.

Casper, J. L. and R. R. Murphy (2002). Workflow study on human-robot interaction in USAR. In *Proc. 2002 IEEE International Conference on Robotics and Automation, May 11-15 2002*, Washington, DC, United States, pp. 1997–2003.

Cassell, J., T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan (1999). Embodiment in conversational interfaces: Rea. *Conference on Human Factors in Computing Systems, May 15-May 20*, 520–527.

Cassell, J., Y. Nakano, T. Bickmore, C. L. Sidner, and C. Rich (2001). Non-verbal cues for discourse structure. In *Association for Computational Linguistics Annual Conference (ACL)*, pp. 106–115.

Cheok, A. D., S. W. Fong, K. H. Goh, X. Yang, W. Liu, F. Farzbiz, and Y. Li (2003). Human Pacman: A mobile entertainment system with ubiquitous computing and tangible interaction over a wide outdoor area. *Mobile HCI*, 209–223.

Cheok, A. D., W. Weihua, X. Yang, S. Prince, F. S. Wan, M. Billinghurst, and H. Kato (2002). Interactive theatre experience in embodied + wearable mixed reality space. *In Proc. International Symposium on Mixed and Augmented Reality, ISMAR*, 59–317.

Chong, N. Y., T. Kotoku, K. Ohba, K. Komoriya, and K. Tanie (2001). Exploring interactive simulator in collaborative multi-site teleoperation. In *Proc. 10th IEEE International Workshop on Robot and Human Communication, Sep 18-21*, Bordeaux-Paris.

Clark, H. H. and S. E. Brennan (1991). Grounding in communication. In J. Resnick, L. Levine and S. Teasley (Eds.), *Perspectives on Socially Shared Cognition*, pp. 127 – 149. Washington D.C.: American Psychological Association.

Clark, H. H. and D. Wilkes-Gibbs (1986). Referring as a collaborative process. *Cognition 22*(1), 1–39.

Cohen, P. R. (1992). The role of natural language in a multimodal interface. In *Proc. of the Fifth Symposium on User Interface Software and Technology*, Monterey, CA, USA, pp. 143 – 149.

Cohen, P. R., M. Dalrymple, F. C. N. Periera, J. W. Sullivan, R. A. Gargan, J. L. Schlossberg, and S. W. Tyler (1989). Synergistic use of direct manipulation and natural language. In *Proc. of the ACM Conference on Human Factors in Computing Systems (CHI 89)*, Austin, Texas, USA, pp. 227 – 233.

Collett, T. H. J. and B. A. MacDonald (2006). Developer oriented visualisation of a robot program. *Proceedings 2006 ACM Conference on Human-Robot Interaction, March 2-4* , 49–56.

Dahlback, N., A. Jonsson, and L. Ahrenberg (1993). Wizard of oz studies: Why and how. In *Workshop on Intelligent User Interfaces*, Orlando, Florida.

Denecke, M. (2002). Rapid prototyping for spoken dialog systems. In *Proceedings of 19th International Conference on Computational Linguistics*, Taipei, Taiwan, pp. 1 – 7.

Drury, J., J. Richer, N. Rackliffe, and M. Goodrich (2006). Comparing situation awareness for two unmanned aerial vehicle human interface approaches. *Proceedings IEEE International Workshop on Safety, Security and Rescue Robotics (SSRR). Gainsburg, MD, USA*, August.

Drury, J. L., H. A. Yanco, and J. Scholtz (2005). Using competitions to study human-robot interaction in urban search and rescue. *Interactions 12*(2), 39–41.

Edan, Y. (1999). Food and agricultural robots. In S. Y. Nof (Ed.), *The Handbook of Industrial Robotics, 2nd Edition*, pp. 1143 – 1155. New York, NY: Wiley.

eMagin (2008). *www.3dvisor.com, accessed June 2008*.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. *Human Factors Society, 32nd Annual Meeting*, 97 – 108.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society 37*, 32–64(33).

Fernandez, V., C. Balaguer, D. Blanco, and M. A. Salichs (2001). Active human-mobile manipulator cooperation through intention recognition. In *2001 IEEE International Conference on Robotics and Automation, May 21-26*, Seoul, South Korea.

Fitzmaurice, G. W. and W. Buxton (1997). An empircal evaluation of graspable user interfaces: Towards specialized, space-multiplexed input. In *Proc. of the Conference on Human Factors in Computing Systems (CHI 97)*, Atlanta, GA, USA, pp. 43 – 55.

Fjeld, M., P. Juchli, and B. Voegtli (2003). Chemistry education: A tangilbe interaction approach. In *Proc. of the INTERACT 03 Conference on Human-Computer Interaction*, Zurich, Switzerland.

Fong, T., C. Kunz, L. M. Hiatt, and M. Bugajska (2006). The human-robot interaction operating system. *Proceedings of 2006 ACM Conference on Human-Robot Interaction, March 2-4*, 41–48.

Fong, T. and I. R. Nourbakhsh (2005). Interaction challenges in human-robot space exploration. *Interactions 12*(2), 42–45.

Fong, T., C. Thorpe, and C. Baur (2002a). Robot as partner: Vehicle teleoperation with collaborative control. *Multi-Robot Systems: From Swarms to Intelligent Automata*, 01 June.

Fong, T., C. Thorpe, and C. Baur (2002b). Robot, asker of questions. In *IROS 2002, Sep 30*, Volume 42 of *Robotics and Autonomous Systems*, Lausanne, Switzerland, pp. 235–243. Elsevier Science B.V.

Fong, T., C. Thorpe, and C. Baur (2003). Multi-robot remote driving with collaborative control. *IEEE Transactions on Industrial Electronics 50*(4), 699–704.

Friedrich, W. (2002). Augmented reality for development, production and service. In *ISMAR 02 International Symposium on Mixed and Augmented Reality*, Darmstadt, Germany.

Fussell, S. R., L. D. Setlock, and R. E. Kraut (2003). Effects of head-mounted and scene-oriented video systems on remote collaboration on physical tasks. In *The CHI 2003 New Horizons Conference Proceedings: Conference on Human Factors in Computing Systems, Apr 5-10*, Ft. Lauderdale, FL, United States.

Gerkey, B., R. Vaughan, and A. Howard (2003). Player/stage project: Tools for multi-robot and distributed sensor systems. In *Proceedings International Conference on Advanced Robotics*, Coimbra, Portugal, pp. 317 – 323.

Giesler, B., T. Salb, P. Steinhaus, and R. Dillmann (2004). Using augmented reality to interact with an autonomous mobile platform. In *Proceedings- 2004 IEEE International Conference on Robotics and Automation*, New Orleans, LA, United States, pp. 1009–1014.

Glassmire, J., M. O'Malley, W. Bluethmann, and R. Ambrose (2004). Cooperative manipulation between humans and teleoperated agents. In *Proceedings*

- 12th International Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems, HAPTICS 2004, Mar 27-28, Chicago, IL, United States.

Goose, S., S. Sudarsky, Z. Xiang, and N. Navab (2003). Speech-enabled augmented reality supporting mobile industrial maintenance. *IEEE Pervasive Computing 2*(1), 65–70.

Greenwald, A. G. (1976). Within subject designs: To user or not to use? *Psychological Bulletin 83*, 314 – 320.

Hauptmann, A. G. (1989). Speech and gestures for graphic image manipulation. *SIGCHI Bull. 20*, 241 – 245.

Horiguchi, Y., T. Sawaragi, and G. Akashi (2000). Naturalistic human-robot collaboration based upon mixed-initiative interactions in teleoperating environment. In *Proc. of the 2000 IEEE International Conference on Systems, Man and Cybernetics, Oct 8-Oct 11*, Nashville, TN, USA.

Huettenrauch, H., K. S. Eklundh, A. Green, and E. A. Topp (2006). Investigating spatial relationships in human-robot interaction. In *Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, Oct 9 - 15*, Beijing, China, pp. 5052 – 5059.

Huettenrauch, H., A. Green, M. Norman, L. Oestreicher, and K. S. Eklundh (2004). Involving users in the design of a mobile office robot. *IEEE Transactions on Systems, Man and Cybernetics, Part C 34*(2), 113–124.

Inagaki, Y., H. Sugie, H. Aisu, S. Ono, and T. Unemi (1995). Behavior-based intention inference for intelligent robots cooperating with human. In *Proceedings of the 1995 IEEE International Conference on Fuzzy Systems, March 20-24*, Yokohama, Japan, pp. 1695 – 1700.

Iossifidis, I., C. Theis, C. Grote, C. Faubel, and G. Schoner (2003). Anthropomorphism as a pervasive design concept for a robotic assistant. In *2003 IEEE/RSJ International Conference on Intelligent Robots and Systems, Oct 27 - 31*, Volume 4, Las Vegas, NV, United States, pp. 3465–3472.

Irawati, S., S. Green, M. Billinghurst, A. Duenser, and H. Ko (2006). An evaluation of an augmented reality multimodal interface using speech and paddle gestures. In *Proceedings of the 16th International Conference on*

Artificial Reality and Telexistence (ICAT 2006), Hangzhou, China, pp. 272 – 283.

Ishii, H. and B. Ullmer (1997). Tangible bits: Towards seamless interfaces between people, bits and atom. In *Proc. of the Conference on Human Factors in Computing Systems (CHI 97)*, Atlanta, GA, USA, pp. 234 – 241.

Ishikawa, N. and K. Suzuki (1997). Development of a human and robot collaborative system for inspecting patrol of nuclear power plants. In *Proceedings of the 1997 6th IEEE International Workshop on Robot and Human Communication, RO-MAN'97, Sep 29-Oct 1*, Sendai, Japan, pp. 118 – 123.

Kanda, T., H. Ishiguro, T. Ono, M. Imai, and R. Nakatsu (2002). Development and evaluation of an interactive humanoid robot Robovie. In *Proc. of the 2002 IEEE International Conference on Robotics and Automation, May 11-15*, Washington, DC, United States, pp. 1848 – 1855.

Kato, H., M. Billinghurst, I. Poupyrev, K. Imamoto, and K. Tachibana (2000). Virtual object manipulation on a table-top AR environment. In *Proc. of the IEEE and ACM International Symposium on Augmented Reality (ISAR 2000)*, Munich, Germany, pp. 111 – 110.

Kato, H., M. Billinghurst, I. Poupyrev, and N. Tetsutani (2001). Tangible augmented reality for human computer interaction. In *Proc. of NICOGRAPH 01*, Nagoya, Japan, pp. 39 – 44.

Kay, P. (1993). Speech-driven graphics: A user interface. *Journal of Microcomputer Applications 16*(3), 223–231.

Kendon, A. (1967). Some functions of gaze direction in social interaction. *Acta Psychologica 32*, 1–25.

Kendon, A. (1983). Gesture and speech: How they interact. *Nonverbal Interaction. J. Wiemann, R. Harrison (Eds). Beverly Hills, Sage Publications*, 13–46.

Kiyokawa, K., M. Billinghurst, S. E. Hayes, A. Gupta, Y. Sannohe, and H. Kato (2002). Communication behaviors of co-located users in collaborative AR interfaces. In *Proc. of the International Symposium on Mixed and Augmented Reality, ISMAR*, Darmstadt, Germany, pp. 139–148.

Kolsch, M., R. Bane, T. Hollerer, and M. Turk (2006). Multimodal interaction with a wearable augmented reality system. *IEEE Computer Graphics and Applications 26*(3), 62–71.

Kuzuoka, H., K. Yamazaki, A. Yamazaki, J. Kosaka, Y. Suga, and C. Heath (2004). Dual ecologies of robot as communication media: Thoughts on coordinating orientations and projectability. In *Proc. of 2004 Conference on Human Factors in Computing Systems, CHI 2004, Apr 24-29*, Vienna, Austria, pp. 65 – 72.

LandTransportNZ (2008). *www.landtransport.govt.nz/roadcode/theory-test-questions/general-questions.html*, accessed October 2008.

Looser, J., R. Grasset, H. Seichter, and M. Billinghurst (2006). Osgart - a pragmatic approach to MR. In *Proc. of the Industrial Workshop at ISMAR 2006*, Santa Barbara, CA, USA.

Maida, J., C. Bowen, and J. Pace (2007). Improving robotic operator performance using augmented reality. In *Proc. of Human Factors and Ergonomics Society 51st Annual Meeting*, Baltimore, Maryland, USA, pp. 1635 – 1639.

Makela, K., E. P. Salonen, M. Turunen, J. Hakulinen, and R. Raisamo (2001). Conducting a wizard of oz experiment on a ubiquitous computing system doorman. In *Proc. of the International Workshop of Information Presentation and Natural Multimodal Dialogue*, Verona, pp. 115 – 119.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago: The University of Chicago Press.

MerriamWebster (2008). *www.merriam-webster.com/dictionary*, accessed September 2008.

MicrosoftSpeech (2007). *www.microsoft.com/speech/, accessed August 2007*.

Milgram, P. and F. Kishino (1994). Taxonomy of mixed reality visual displays. *IEICE Transactions on Information and Systems E77-D*(12), 1321–1329.

Milgram, P., S. Zhai, D. Drascic, and J. Grodski (1993). Applications of augmented reality for human-robot communication. In *Proceedings of IROS 93: International Conference on Intelligent Robots and Systems*, Yokohama, Japan, pp. 1467–1472.

Minneman, S. and S. Harrison (1996). A bike in hand: A study of 3D objects in design. In N. Cross, H. Christiaans, and K. Dorst (Eds.), *Analyzing Design Activity*. Chichester: J. Wiley.

Morita, T., K. Shibuya, and S. Sugano (1998). Design and control of mobile manipulation system for human symbiotic humanoid: Hadaly-2. In *Proceedings of the 1998 IEEE International Conference on Robotics and Automation, May 16-20*, Leuven, Belgium, pp. 1315 – 1320.

Murphy, R. R. (2004). Human-robot interaction in rescue robotics. *IEEE Transactions on Systems, Man and Cybernetics, Part C 34*(2), 138–153.

Nass, C., J. Steuer, and E. R. Tauber (1994). Computers are social actors. In *Proceedings of the CHI'94 Conference on Human Factors in Computing Systems, Apr 24-28*, Boston, MA, USA, pp. 72–78.

Nielsen, J. (1994). In *Usability Engineering*. San Francisco, CA: Morgan Kaufmann, Inc.

Nikishkov, G. and T. Tsuchimoto (2007). Using augmented reality for real-time visualization of tactile health examination. In *Proc. of the GRAPP 07 Conference on Computer Graphics Theory and Applications*, Barcelona, Spain, pp. 91 – 97.

Nilsen, T., S. Linton, and J. Looser (2004). Motivations for AR gaming. In *Proc. of the Fuse 04 New Zealand Game Developers Conference*, Dunedin, New Zealand, pp. 86 – 93.

Nintendo (2008). *http://www.nintendo.com/wii, accessed July 2008*.

NIST (2008). *e-HandBook of Statistical Methods, www.itl.nist.gov/div898/handbook/prc/section4/prc47.htm, accessed September 2008*.

Nourbakhsh, I. R., J. Bobenage, S. Grange, R. Lutz, R. Meyer, and A. Soto (1999). Affective mobile robot educator with a full-time job. *Artificial Intelligence 114*(1-2), 95–124.

Nourbakhsh, I. R., K. Sycara, M. Koes, M. Yong, M. Lewis, and S. Burion (2005). Human-robot teaming for search and rescue. *IEEE Pervasive Computing 4*(1), 72–77.

NXT++ (2007). *www.nxtpp.sourceforge.net/, accessed August 2007*.

Ohba, K., S. Kawabata, N. Y. Chong, K. Komoriya, T. Matsumaru, N. Matsuhira, K. Takase, and K. Tanie (1999). Remote collaboration through time delay in multiple teleoperation. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'99): Human and Environment Friendly Robots with High Intelligence and Emotional Quotients', Oct 17-Oct 21*, Kyongju, South Korea, pp. 1866 – 1871.

OpenSceneGraph (2008). *www.openscenegraph.org, accessed June 2008*.

Oviatt, S. (2003). Advances in robust multimodal interface design. *IEEE Computer Graphics and Applications 23*(5), 62–68.

Oviatt, S. L. (2000). Designing the user interface for multimodal speech and gesture applications: State of the art systems and research directions. *Human Computer Interaction 15*(4), 263 – 322.

Perzanowski, D., D. Brock, S. Blisard, W. Adams, M. Bugajska, A. Schultz, G. Trafton, and M. Skubric (2003). Finding foo: A pilot study for a multimodal interface. In *Proc. of the IEEE System, Man, and Cybernetics Conference*, Washington, DC, USA, pp. 3218 – 3223.

Piekarski, W. and B. Thomas (2002). Arquake: The outdoor augmented reality gaming system. *Communications of the ACM 45*(1), 36 – 38.

Prince, S., A. D. Cheok, F. Farbiz, T. Williamson, N. Johnson, M. Billinghurst, and H. Kato (2002a). 3-D live: Real time captured content for mixed reality. In *Proceedings of the International Symposium on Mixed and Augmented Reality, ISMAR2002, Sept 30 - Oct 1*, pp. 7 – 13.

Prince, S., A. D. Cheok, F. Farbiz, T. Williamson, N. Johnson, M. Billinghurst, and H. Kato (2002b). 3-D live: Real time interaction for mixed reality. In *Proc. of the Eighth Conference on Computer Supported Cooperative Work (CSCW 2002), Nov 16-20*, New Orleans, LA, United States, pp. 364 – 371.

Ralph, M. and M. Moussa (2005). Human-robot interaction for robotic grasping: A pilot study. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and System (IROS 2005)*, Edmonton, Alberta, Canada, pp. 454 – 459.

Rani, P., N. Sarkar, C. A. Smith, and L. D. Kirby (2004). Anxiety detecting robotic system - towards implicit human-robot collaboration. *Robotica 22*(1), 85–95.

Reitmayr, G. and D. Schmalstieg (2004). Collaborative augmented reality for outdoor navigation and information browsing. *In Proc. Symposium Location Based Services and TeleCartography 2004 Geowissenschaftliche Mitteilungen Nr. 66*.

Roy, D., K.-Y. Hsiao, and N. Mavridis (2004). Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man and Cybernetics, Part B 34*(3), 1374–1383.

Salber, D. and J. Coutaz (1993). Applying the Wizard of Oz technique to the study of multimodal systems. In *Proceedings of EWHCI/93*, Moscow, Russia, pp. 219 – 230.

Sareika, M. and D. Schmalstieg (2007). Urban sketcher: Mixed reality on site for urban planning and architecture. *In Proc. of the 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, 27 – 30.

Schmandt, C., M. S. Ackerman, and D. Hindus (1990). Augmenting a window system with speech input. *Compuer 23*(8), 50 – 56.

Scholtz, J. (2002). Human robot interactions: Creating synergistic cyber forces. *In A. Schultz and L. Parker, eds., Multi-robot Systems: From Swarms to Intelligent Automata, Kluwer*.

Scholtz, J. (2003). Theory and evaluation of human robot interactions. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, pp. 1 – 10.

Scholtz, J., B. Antonishek, and J. Young (2005). A comparison of situation awareness techniques for human-robot interaction in urban search and rescue. In *Proc. of the CHI 2005, April 2- 7*, Portland, Oregon, USA, pp. 2192 – 2201.

Shelton, B. and N. Hedley (2002). Using augmented reality for teaching earth-sun relationships to undergraduate geography students. In *Proc. of the First IEEE International Augmented Reality Toolkit Workshop*, Darmstadt, Germany, pp. 1 – 8.

Sidner, C. L. and C. Lee (2003). Engagement rules for human-robot collaborative interactions. In *System Security and Assurance, Oct 5-8*, Volume 4 of *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, Washington, DC, United States, pp. 3957–3962.

Sidner, C. L. and C. Lee (2005). Robots as laboratory hosts. *Interactions 12*(2), 24–26.

Siltanen, S., M. Hakkarainen, O. Korkalo, T. Salonen, J. Saaski, C. Woodward, T. Kannetis, M. Perakakis, and A. Potamianos (2007). Multimodal user interface for augmented assembly. In *Proc. of the IEEE 9th Workshop on Multimedia Signal Processing*, Crete, Greece, pp. 78 – 81.

Skubic, M., D. Perzanowski, S. Blisard, A. Schultz, W. Adams, M. Bugajska, and D. Brock (2004). Spatial language for human-robot dialogs. *IEEE Transactions on Systems, Man and Cybernetics, Part C 34*(2), 154–167.

Skubic, M., D. Perzanowski, A. Schultz, and W. Adams (2002). Using spatial language in a human-robot dialog. In *Proc. of the 2002 IEEE International Conference on Robotics and Automation, May 11-15*, Washington, DC, United States, pp. 4143 – 4148.

Soler, N. L., J. Schmid, C. Koehl, J. Marescaux, X. Pennec, and N. Ayache (2004). Virtual reality and augmented reality in digestive surgery. In *Proc. of the ISMAR 04 International Symposium on Mixed and Augmented Reality*, Arlington, VA, USA, pp. 278 – 279.

Tenbrink, T., K. Fischer, and R. Moratz (2002). Spatial strategies in human-robot communication. *Korrekturabzug Kuenstliche Intelligenz, Heft 4/02, pp 19-23, ISSN 0933-1875, arendtap Verla, Bemen.*

TheLegoGroup (2007). *mindstorms.lego.com/, accessed August 2007.*

Thrun, S. (2004). Toward a framework for human-robot interaction. *Human-Computer Interaction 19*, 9–24.

Trochim, W. M. K. (2006). Research methods knowledge base. *www.socialresearchmethods.net/kb/index.php.*

Tsoukalas, L. H. and D. T. Bargiotas (1996). Modeling instructible robots for waste disposal applications. In *Proceedings of the 1996 IEEE International Joint Symposia on Intelligence and Systems, Nov 4-5*, pp. 202 – 207.

Tversky, B., P. Lee, and S. Mainwaring (1999). Why do speakers mix perspectives? *Spatial Cognition Computing 1*, 399–412.

Watanuki, K., K. Sakamoto, and F. Togawa (1995). Multimodal interaction in human communication. *IEICE Transactions on Information and Systems E78-D*(6), 609–615.

Yanco, H. A. and J. L. Drury (October 2004). Where am I? acquiring situation awareness using a remote robot platform. In *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, The Hauge, Netherlands, pp. 2835 – 2840.

Yanco, H. A., J. L. Drury, and J. Scholtz (2004). Beyond usability evaluation: Analysis of human-robot interaction at a major robotics competition. *Human-Computer Interaction Human-Robot Interaction 19*(1-2), 117–149.

ZeroC (2008). *www.zeroc.com/*, accessed June 2008.